Scoring Story Recall for Individual Differences Research: Central Details, Peripheral Details, and Automated Scoring

David Martinez

Applied Research Lab for Intelligence and Security (ARLIS),

University of Maryland, College Park

Author Note

Contact Information

David Martinez, dmartin5@umd.edu, https://orcid.org/0000-0001-8265-3862

Data, Code, and other Materials

Materials, data, and scripts are available at https://osf.io/5qxkh/

Funding

Funding for this work was provided by an Internal Research and Development grant from ARLIS as well as a grant from the Office of Naval Research (N00014-21-1-2220).

Acknowledgments

I would like to thank Abigail Rosenberg, Ashley Higgins, Christina Yang, Emily Blume, Esha Edakoth, Jarrett Lee, Maia Brandolini, Margaret Kimbis, Mitch Koff, Ramata Lam, Riki Engling, and Shawna Dougherty for their help throughout the project.

Abstract

Story recall is a popular declarative memory paradigm among clinical and geropsychologists. It is less popular among individual differences researchers. One reason differential psychologists may favor other paradigms is that the prospect of scoring story recall is daunting, as it typically requires the subjective scoring of hundreds or thousands of freely recalled narratives. In this study, I investigated two questions related to scoring story recall. First, whether there is anything to gain by scoring story recall for memory of central and peripheral details or if a single score is sufficient. Second, I investigated whether scoring can be automated using computational methods—namely, BERTScore and GPT-4. A total of 235 individuals participated in this study. Results suggest that memory for central and peripheral details largely rely on the same cognitive processes, rendering the differentiation between central and peripheral details redundant. Regarding automated scoring, both BERTScore and GPT-4 derived scores were strongly correlated with manually derived scores; additionally, factors estimated from the various scoring methods all showed a similar pattern of correlations with fluid intelligence, crystallized intelligence, and working memory capacity. Thus, researchers may be able to streamline scoring by disregarding detail type and by using automated approaches. Further research is needed, particularly of the automated approaches, as both BERTScore and GPT-4 derived scores were occasionally leptokurtic.

Keywords: story recall; prose recall; individual differences; automated scoring; long-term memory

Introduction

Story (or prose) recall is a fairly popular declarative memory paradigm among clinical psychologists and those interested in the effects of aging (e.g., Baek et al., 2011; Ratner et al., 1987; Ryan & Gontkovsky, 2023). In general, in a story recall task, individuals listen to or read a short story (~70 words) and then recall the story as well as possible, either immediately after presentation or a delay. Perhaps the most widely used story recall task is the Logical Memory subtest of the Wechsler Memory Scale (Cassady & Chittooran, 2010; Wechsler, 2009). Such tests appear to be sensitive indicators of memory decline as associated with typical aging but also mild cognitive impairment and Alzheimer's disease (Rabin et al., 2009; Ratner et al., 1987; Robinson-Whelen & Storandt, 1992), hence their popularity amongst clinical and geropsycholgists.

Interestingly, story recall does not appear to be as popular among researchers interested in individual differences in neurotypical young adults. Story recall is absent from a recent review of individual differences in long-term memory (Unsworth, 2019) as well as a meta-analysis investigating memory and creativity (Gerver et al., 2023). Instead of story recall, differential psychologists tend to prefer tasks such as free recall of lists, cued recall of paired-associates, and various recognition tasks (see Gerver et al., 2023; Unsworth, 2019).

There are number of reasons differential psychologist might opt for another declarative memory paradigm over story recall. For example, controlling for a variety of linguistic and other features will be easier when one can place words into arbitrary lists, rather than in coherent stories. Another possible reason for this preference is that, compared to story recall, other measures are easier to score.

The prospect of scoring story recall for an individual differences study is daunting. To achieve appropriate statistical power and precision, differential psychologists would likely wish to administer more than one story to hundreds of participants. Each response will have to be scored in, ideally, as objective a manner as possible. This typically means that researchers will manually score hundreds of

responses based on some list of criteria. Additionally, questions about what and how to score will inevitably surface. For example, should one score for gist recall or for recall of all details?

Despite these difficulties, there are reasons differential psychologists may wish to include story recall tasks in their studies. First, story recall mimics the kind of content we often recall in our lives. While it is certainly true that in industrialized societies we often memorize lists and pairs of words (cf. Mandler et al., 1980), we also often recount narrative episodes from our day-to-day lives. Thus, story recall exhibits ecological validity. Second, because stories are cohesive narratives, they may place more emphasis on experiential and linguistic factors and their integration in working memory (Baddeley & Wilson, 2002; Jefferies et al., 2004). As such, story recall can be used to investigate factors that are often controlled in individual differences research. Relatedly, story recall can provide evidence of convergent validity. We should expect that story recall, free recall, and paired-associate learning exhibit similar relationships with other variables. When that is not the case, we should be able to provide an explanation for this difference. For example, as mentioned above, there is reason to expect that story recall would be more strongly correlated with one's store of linguistic and general knowledge; therefore, we may see larger correlations between verbal ability and story recall than is seen with other declarative memory tasks. In sum, we can learn more about human memory by including story recall in our arsenal of long-term memory tasks.

Current Study

The general purpose of this study was to investigate 1) whether points should be awarded separately for central and peripheral details or as a single total score; and 2) whether computational methods can be used to automatically score story recall. To evaluate these scoring methods, I investigated correlations between the methods and with variables known to be related to declarative long-term memory, namely: fluid intelligence, crystallized intelligence, and working memory capacity (Unsworth, 2019).

How should points be awarded?

The concern here was whether recalled story details should be summed to form a single score (e.g., Wechsler, 1997) or whether recall of *central* and *peripheral* details should form two separate scores (e.g., Sacripante et al., 2023; see also Wechsler, 2009; Dunn et al., 2002). In story recall, responses are typically scored based on the number of details recalled. These details can be classified as central or peripheral details (e.g., Sacripante et al., 2023). Central details are related to the *gist* of a story (i.e., the theme and plot) and tend to be abstracted; peripheral details are all other details in a story, are more specific than central details, and tend to be more numerous. For example, consider a story about a father and son who crashed into a tree while riding a bicycle (Sacripante et al., 2023). The central details may be that a parent and child were involved in an accident; the peripheral details may include the gendered relationship (father/son), the mode of transportation, and what they struck.

Research and theory suggest that cognitive processes supporting memory for central and peripheral details differ (Brainerd & Reyna, 2001; Sacripante et al., 2023). For instance, central details appear to be easier to learn and are forgotten at a slower rate compared to peripheral details (Brainerd & Reyna, 1990; Foldi, 2011; Glenberg et al., 1987; Sacripante et al., 2023). This may be because as individuals read prose passages, they are retrieving and integrating prior knowledge as well as making inferences (Allen & Baddeley, 2009; Baddeley et al., 2009; Jefferies et al., 2004). It is the central details that likely receive this extra processing which may lead to enhanced retrievability. Relatedly, central details, such as the main character in a story, are often elaborated on by the peripheral details. Elaboration is known to benefit memory (e.g., Bartsch & Oberauer, 2021), as is the repeated retrieval one likely engages in when central details are referenced throughout a narrative (Roediger & Butler, 2011; Rowland, 2014).

Still, it is possible that there is a single memory system that supports memory for central and peripheral details. For example, it may be that central details *do* receive increased processing, which aids

memory, but the same processes operate on peripheral details. Additionally, one should expect a significant correlation between memory for central and peripheral details because if one can recall a peripheral detail, then one can likely recall associated central details (for similar arguments, see Abikoff et al., 1987; Dunn et al., 2002). In the story about the bicycle accident above, if one recalls that a tree was struck, then one would almost certainly recall that there was an accident.

If memory for central and peripheral details differ, then one can maximize information gain by scoring responses for *both* central and peripheral details. If, on the other hand, memory for peripheral and central details rely on similar cognitive processes, then one can improve reliability by generating a single *total* score rather than splitting into two scores.

In this study, details were classified as either central or peripheral in a similar fashion as done by Sacripante and colleagues (2023). To evaluate the dissociability of memory for central and peripheral details, I analyzed correlations between the scoring methods as well as with other variables. A perfect correlation between variables representing the two types of memory would suggest that they are one and the same. Anything less than a perfect correlation could be because of measurement error or because, in fact, the two variables are assessing different processes. In the case of a less than perfect correlation, it is helpful to look at correlations with other variables; if the two memory factors correlate differently with other variables, then the less than perfect relationship is less likely due to measurement error. Here, I investigated correlations between the memory factors and fluid intelligence, crystallized intelligence, and working memory—factors known to be related to declarative memory performance (for a review, see Unsworth, 2019).

Can automated scoring approaches perform as well or better than humans?

The second goal was to examine whether computational methods can be used to automatically score story recall. Clearly, an automated approach would reduce the amount of time and effort needed to score hundreds (or thousands) of responses, but it can also potentially increase the reliability and

validity of story recall. As argued by Chandler and colleagues (2019, 2021), a computational approach to scoring story recall can provide a more accurate and continuous measure of recall from semantic memory. For example, in the bicycle story (above), an individual may recall that the bicycle *slammed* into a tree rather than *crashed* into it. Depending on one's scoring scheme, one may award 0, .5, or 1 point for recalling this detail (e.g., Abikoff et al., 1987; Power et al., 1979). Using a computational approach, one can instead award points in a more objective manner based on how semantically similar *crashed* and *slammed* are to each other—in this case, using the language model and semantic space used in this study (see Methods), one would award .8348 points. The objective nature of automated scoring along with its precision should increase reliability. Validity is increased as a consequence of the increase in reliability and because the score more accurately captures the fidelity of the recalled memory.

While computational methods have been used in the past (e.g., Chandler et al., 2019, 2021; Dunn et al., 2002) to automate story recall scoring, 1) it has been rare for researchers to make their materials freely available and 2) computational methods continue to advance. Chandler et al. (2021), for example, used BERTScore (Zhang et al., 2020) to assess the semantic similarity between target prose passages and participant responses. BERTScore is a metric that uses embeddings (numerical representations of words and their context) derived from language models, such as BERT, a natural language processing model trained on Wikipedia and Google's book corpus (Devlin et al., 2019). Chandler and colleagues (2021) reported a correlation of .86 between human-derived scores and those derived using BERTScore. Unfortunately, the materials are not freely available, and so it is difficult to replicate and build upon this work. Additionally, more advanced models are now available. GPT-3 and GPT-4 (OpenAI, 2022;2023) are large language models that have been trained on larger corpora than previous language models and appear to be outperforming humans as well as older language models on a number of tasks (e.g., T. Brown et al., 2020; Strong et al., 2023).

Method

Participants

Participants from this study were drawn from a larger effort examining individual differences in creative thinking. After completing three 2-hour sessions for the creative thinking project, participants were invited to participate in a 30 min study examining memory. All participants were recruited from the online recruitment platform Prolific or through the psychology research participant pool at the University of Maryland. Eligibility criteria are listed below in Table 1.

Table 1

Eligibility Criteria

General Eligibility	Language History Eligibility
• Age between 18 and 35 • Normal or corrected-to-normal vision • Normal hearing • Unimpaired use of dominant hand • Must be residing in the United States.	• Native English speaker or learned American English by age 6 • Dominant (strongest) language is American English • Began attending school in the United States by age 6 • Has lived in the United States continuously since (at least) age 6
Technological and Other Requirements	
• Desktop or personal laptop with keyboard, mouse, and camera • Windows or Mac operating systems • Private, distraction-free space to complete the study • Sufficiently fast (e.g., 1 mbps download/upload speed) and steady internet connection • Microphone or be able to call in to the teleconferencing platform • Must have the ability and be willing to download free software (e.g., Zoom, Google Chrome)	

Prolific participants were compensated with approximately $80 for completing the creative thinking portion of the study and approximately $10 for completing the story recall tasks. Maryland students were compensated with 1 course credit or $15 per hour of participation. In accordance with the University of Maryland Institutional Review Board, informed consent was always obtained prior to participation.

Sample size was determined based on a combination of heuristics and resource availability. Based on prior work investigating individual differences in declarative long-term memory (e.g., Martinez & Singleton, 2019; Wilhelm et al., 2013), we estimated that we would need up to 230 participants to reliably detect latent variable correlations between the declarative memory factor(s) and fluid intelligence, crystallized intelligence, and working memory capacity. Given the novelty of other aspects of this study, it was difficult to estimate how many participants would be needed to, for example, detect meaningful differences amongst different scoring methods. In total, 249 individuals expressed interest in completing this study, however only 235 participants did so. Demographic information is provided in Table 2.

Procedure

The study consisted of a pre-screen, three online proctored sessions, and an online unproctored session. The pre-screen was administered as an unproctored survey to Prolific participants; research pool participants completed the pre-screen at the beginning of Session 1. During the pre-screen (~ 15 min), participants learned about the study goals, consented to participate, and completed a demographic questionnaire.

Sessions 1, 2, and 3 were proctored. These sessions were hosted on the teleconferencing platform Zoom. Participants were asked to share their screens and to always keep their camera and microphone on except during breaks. There was always a research assistant monitoring participants as they completed a battery of tasks assessing a variety of cognitive abilities. All tasks used in the current study are displayed in Table 3 and a full list of tasks used in the creativity study is available at https://osf.io/5njbs/. These first three sessions were estimated to take no longer than 2 hours.

After completing session 3, participants were invited to participate in an unproctored study investigating memory. Participants were asked to complete the session in a quiet, distraction free environment and on laptops or desktops. The session was estimated to take approximately 30 min.

Table 2

Demographics

Age Range: 18-35 M = 27.18 SD = 4.99	**Hispanic/Latino? (%)** Yes: 13.3 No: 85.8 Decline to state: 0.9	**US Census "Race" Category (%)** American Indian or Alaskan Native: .4 Asian: 11.2 Black: 10.3 White: 70 Other/Mixed Race: 7.3 Decline to State: 0.9
Sex (%) Male: 46.8 Female: 49.8 Non-Binary: 2.6 Prefer not to say: 0.9	**Household Income (%)** < $30,000: 12.9 $30,000-$59,999: 19.7 $60,000-$89,999: 27 $90,000-$119,999: 16.3 > $120,000: 24	**Highest Level of Education (%)** High School Diploma: 9 Some college but did not/have not graduated: 30 Associate degree: 7.3 Bachelor's Degree: 33 Master's Degree: 16.7 Other Professional Degree: 1.7 PhD, MD, or JD: 2.1
US Region (%) West: 23.2 Midwest: 23.2 Northeast: 40.3 South: 12.4 Missing: 0.9	**Current Student? (%)** Yes: 32.6 No: 67.4	

Table 3

Tasks by Sessions

Pre-screen	Session 1	Session 2	Session 3	Session 4
Consent Demographic Questionnaire	Re-consent LNS Ravens Vocabulary Information RACO	Re-consent Number Series Pattern Span Matchstick Math Grammar	Re-consent RMS Verbal Analogies	Re-consent Story Task 1 Story Task 2 Story Task 3

Note. LNS = letter-number sequencing; RACO = route and count; RMS = running memory span

Tasks

All cognitive tasks were administered online through a proprietary platform. The majority of tasks began with audiovisual instructions, a demonstration of the task, and an example item. For some tasks, we suggested participants use particular strategies. For example, we encouraged participants to use the constructive-matching strategy (Bethell-Fox et al., 1984; Gonthier & Thomassin, 2015) when

completing fluid intelligence items. This was done to reduce measurement noise owing to individual differences in strategy use and experience with specific tasks. Because our tasks were administered via our stimulus administration platform, the tasks are unavailable; however, to aid replication efforts, videos demonstrating our tasks and stimuli are available at https://osf.io/5qxkh/.

Story recall

Stories were drawn from Chandler et al. (2019, 2021), Cunje et al. (2007), Holmlund et al. (2020), Sacripante et al. (2023), and Schnabel (2012). Minor edits were made to some of the stories, for example, replacing proper nouns with more generic terms—edits were generally made to improve readability and to distinguish between central and peripheral details more clearly. In general, stories tended to be short. The stories used in Story Task 1 and Story Task 2 ranged between 62 and 77 words; the stories used in Story Recall 3 were all from Cunje et al. (2007) and individually were 30 or 29 words long (including titles), however, in this case all three stories were administered at once for a total word count of 89 words (see below for more details). Additionally, all stories employed topics and vocabulary that should be familiar to most adults. Readers interested in specific lexical characteristics of these stories can refer to the works cited above.

All story recall tasks contained three stories. Participants were told that they should try to remember as much of the stories as possible and, when asked to recall, if they forgot details, they should try to continue with whatever information they can recall.

Story Task 1. In this task, three stories were presented visually as text on the computer screen one at a time, for 45 s each. Based on internal pilot studies, this amount of time was deemed sufficient for the vast majority of participants to finish reading each story. Participants were not told what they should do if they finished reading the story in under 45 s, however, they were not allowed to move on until after 45 s. After each story was presented, participants were asked to read five sentences and make grammaticality judgments (e.g., “During the week of final spaghetti, I felt like I was losing my mind”;

Kane et al 2004). The purpose of the filler task was to distract and interrupt rehearsal. After making grammaticality judgments, participants had up to two minutes to recall as much of a story as possible. This procedure was repeated for the next two stories.

Story Task 2. This task was very similar to Story Task 1, but stories were presented aurally, and the filler task was equation verification. The stories were read by the present author. The audio clips were approximately 21 s, 22 s, and 25 s long. The filler task began immediately after the audio clip ended; thus, participants had a single chance to hear each story. The equation verification task—e.g., (4x4) + 1 = 17—was drawn from Kane et al. (2004). There were five equations between presentation and recall of any story.

Story Task 3. This task employed three very short (~30 words each), structurally and thematically similar stories developed by Cunje et al. (2007), and no filler task. All stories were presented at once and participants were given 90 s to read and memorize all three stories. Immediately after 90 s, participants were shown three text boxes with each identified by the title of one of the three stories they had just read. They then had 90 s to try to recall all three stories. It was expected that difficulty would be increased by the interference resulting from the similarity across stories. For example, in one story, the setting was described as “a hot May Morning” and in another, it was described as, “a warm September afternoon.”

Fluid intelligence (Gf)

Fluid intelligence was estimated using Ravens matrices (Raven et al., 1998), number series (Thurstone, 1938), verbal analogies (largely drawn from Kane et al., 2004 as well as online sources), and matchstick math (inspired by Knoblich et al., 1999). The first three tasks are inductive reasoning tasks. Such tasks are often used to assess fluid intelligence. Matchstick math was added after an initial attempt at analyzing the data resulted in a non-positive definite covariance matrix. Matchstick math is perhaps more properly considered a measure of creative problem-solving, however, our data indicated that it is

strongly correlated with the other measures of fluid intelligence and could be used to aid in estimating fluid intelligence.

Ravens. Participants attempted to complete 18 odd items from Ravens Advanced Progressive Matrices (Raven et al., 1998) in 10 minutes.

Number series (Thurstone, 1938). In this task, participants were presented with a sequence of numbers. Their task was to identify which number, from presented options, continued the sequence. There were 15 items and a 5 minute time limit.

Verbal analogies. Verbal analogy items were largely drawn from Kane et al. 2004 and supplemented with verbal analogies from various online sources. Participants attempted to complete 18 items within 6 minutes.

Matchstick math. In matchstick, participants were presented with incorrect mathematical statements formed from matchsticks. In contrast to Knoblich et al. (1999) who presented Roman numerals, we used Arabic numerals. Participants were to move a single matchstick to correct the mathematical statement. For example, 3 + 5 = 6 can be corrected by moving a single matchstick (here represented by the vertical or horizontal segments). There were two parts to this task, and each part had 5 items. In part one, participants were given 2 minutes and 45 seconds to complete each item. If participants did not submit a response after 1 minute and 30 seconds, then they were provided with a hint; after 2 minutes and 15 seconds another hint was presented; after 2 minutes and 45 seconds, the solution was revealed. In part two, participants were given 2 minutes to complete each item and received a single hint after 1 minute 30 seconds; solutions were not provided. Scores were calculated by assigning a maximum of 3 points for each item in part 1 and 2 points for each item in part 2. Points were deducted from the maximum for each hint a participant observed before solving an item. The maximum score was 25.

Crystallized Intelligence (Gc)

Crystallized intelligence was assessed with two verbal ability tasks (vocabulary and grammar) and a general knowledge test. All tasks used a multiple-choice response format.

Vocabulary (Ekstrom et al., 1976). There are two sections for this task, each with 24 items with a 6-minute time limit. Each item consists of a target word and 5 options that could potentially mean the same thing or nearly the same thing as the target word. Participants are instructed to select the answer choice they believe to mean the same thing or nearly the same thing as the target word.

Grammar (Martinez & Singleton, 2019). The test consists of 21 "improving sentences" items selected from official SAT practice tests released between 2004 and 2013. Items were selected such that they ranged in difficulty and were generally organized from easiest to most difficult. Each item consists of a sentence with a portion underlined; the participant was to select the answer choice that best rephrased the underlined portion or, if the original phrasing was the best choice, select the first answer choice, which always repeated the original phrasing. Participants had up to 10 min to complete the test.

Information. The test consists of 40 general knowledge questions. Participants are given 7 minutes to complete the task. Questions are on a variety of topics, though largely material one would learn in school.

Working memory capacity

Working memory capacity was estimated using immediate memory tasks that disrupt or otherwise impede rehearsal. Initially, I planned to estimate working memory capacity using a total of four tasks, however, one task—running letter span (Bunting et al., 2006)—was weakly correlated with our visuospatial working memory tasks ($r < .15$). Consequently, running letter span was removed from analyses.

Letter-Number Sequencing (Mielicki et al., 2018). In the letter-number-sequence task, numbers and letters flash onto the screen one at a time and the participant's task is to remember the numbers

and letters and arrange them in sequence. The participant must recall the numbers first in numerical order followed by the letters in alphabetical order. The participant receives points for each number and letter they recall in the correct position. There were 18 critical items, with three items each at set sizes 4-9.

Route and Count. This task is a test of visual working memory inspired by the Remember and Count task (Hughes et al., 2016). This task is a Brown-Peterson task (J. Brown, 1958; Peterson & Peterson, 1959) consisting of a block tapping task (Kessels et al., 2000; Milner, 1971) followed by a distracting counting task (taken from Engle et al., 1999). In this task, there are nine rectangles presented on screen; two to five of the nine rectangles flash in a sequence, with each rectangle flashing for 1 s. After the sequence of squares, an image of teal and dark blue circles and rectangles appear on the screen as a distraction, and participants are required to count the number of dark blue circles on the screen. Following this, the original image of the 9 rectangles appears on the screen, and participants recall the order of the two to five flashing rectangles by clicking on the rectangles in the order they were presented. There was a total of 12 trials, with three trials each at set sizes 2-5. Points were received for correctly recalling rectangles in their serial position; no penalty was incurred for incorrectly counting the circles in the distracting task.

Pattern Span (Della Sala et al., 1999; Martinez & Singleton, 2018). Participants see 5x6 grids with white and black squares forming an abstract pattern for 3 s. Immediately following the encoding period, a screen with black and white static visual noise is presented for 300 ms, followed by an empty grid. The participant is to reproduce the pattern they were previously shown. There were 18 items. Participants received one point for every item correctly recalled—that is, item scoring was all-or-nothing.

Story Recall Scoring

Manual Scoring

Story recall tasks were scored by five trained raters. The document used to train raters—which includes the criteria we used—is available at https://osf.io/5qxkh/. The author and another researcher deconstructed each story into *conceptual units* largely composed of noun phrases (e.g., "riding a bike") but also individual words that the author and research assistant *subjectively* considered significant (e.g., “crashed”). These conceptual units are also provided in the project repository. Importantly, note that after analyzing the data, I noticed some peripheral details were highly similar to central details and so I removed these redundant peripheral details and scores were recalculated. This was done on the grounds that it would be uninteresting to find a large correlation between factors representing memory for peripheral and central details if there was a large degree in overlap in the details used to estimate the two factors. This change did not alter the results of the study.

Participants were awarded one point for each conceptual unit recalled verbatim or if it was deemed highly similar; half a point was awarded when a conceptual unit was deemed to be similar but not similar enough to merit a full point; no points were awarded otherwise. As an example, in the complete bicycle story, at some point the father tries to brake with his shoes. “Dad tries to break with his shoes” was a conceptual unit and a response that would receive one full point was, “dad tried to break with his foot”. Although *shoe* was replaced by *foot,* the inferred meaning is nearly identical.

Despite the subjectivity of this scoring method, interrater reliability was strong. Raters first scored a subset of nearly 15% of all responses. Responses for this reliability assessment were drawn pseudo-randomly from across all nine stories such that there was an equal number of responses per story. In line with prior research using story recall, interrater reliability was quite strong (Abikoff et al., 1987; Holmlund et al., 2020; Power et al., 1979). At the level of the conceptual unit, Pearson correlations

amongst the five raters ranged between .72 and .97 per story; when considering total scores, Pearson correlations ranged between .83 and .98 per story. Given this level of interrater reliability, the remaining ~85% of responses were divided equally among raters (i.e., each remaining response was scored by one of five raters). Those responses that were scored by multiple raters were averaged across raters and included in analyses.

Scoring with BERTScore

First, participants' responses were corrected using the spell check function within Microsoft Excel. No other corrections (e.g., for grammar) were made. Next, all responses were scored using the default BERTScore settings (Zhang et al., 2020) and microsoft/deberta-xlarge-mnli model (He et al., 2021), as currently recommended by the authors of BERTScore. Participant responses were set as the candidate sentences and the target stories were set as the references. The recommended F1 score was used to index participant performance (see Zhang et al., 2020)—an F1 score ranges between zero and one; the higher the score, the closer two texts are in meaning.

Scoring with GPT-4

The chat completions function of OpenAI's GPT-4 application programming interface (API) was used to score responses. Participant responses were provided to the GPT models as is—that is, neither spelling or grammar was corrected, nor any other modifications made. First, I piloted different instructions (prompts) to give GPT-4 using the GPT-3.5 chat completions API as well as ChatGPT with GPT-3.5. Initial pilots with about 10 percent of the data included simply instructing GPT-3.5 to compare participant responses with target stories and provide a semantic similarity score. I then expanded on these instructions by asking it to rate similarity on a scale of 1-10, 1-20, and 1-100. Performance appeared to be highly skewed and scores were always whole numbers. Next, I provided instructions that were similar to the instructions provided to the human raters and asked it to compare participant

responses to the list of details for each story—giving 0, .5, or 1 point for each detail recalled, depending on semantic similarity—sum the points, and provide a single total score per response. Although, there were occasional errors in which scores were given per detail (rather than a total score) or GPT-3.5 would provide a rationale in addition to the total score, the scores appeared more normally distributed. Once the prompt was finalized, all responses, instructions and details were submitted to GPT-4 to score the entire set of responses. Additionally, the temperature parameter was set to zero to make the model output more deterministic and all other defaults were used. Finally, as with the manually scored responses, redundant details were removed from the list of details.

Analysis packages

Data were gathered and aggregated in R (R Core Team, 2021) using the Psych (Revelle, 2017), plyr (Wickham, 2011), and dplyr (Wickham et al., 2015) packages. Visualizations were created using ggplot2 (Wickham, 2016). Latent variable analyses were conducted using lavaan (Rosseel, 2012) while semplot (Epskamp, 2015) was used to aid in visualizing latent variable models. The reticulate package (Ushey et al., 2020) was used with BertScore (Zhang et al., 2020). The httr (Wickham, 2020) and jsonlite (Ooms, 2014) packages were used to interface with OpenAI's GPT models.

Analyses and Results

Descriptive Statistics

Given the large number of variables and the focus of this article, individual task descriptives (including reliability estimates) and bivariate correlations are relegated to the Appendix and repository (https://osf.io/5qxkh/). Some descriptive statistics and correlations specific to the story recall tasks are provided farther below in Table 6 under the section entitled "Further Investigations of the Adequacy of Automated Scoring Methods".

Should Scores be Tabulated by Detail Type?

Differences in Recall by Type of Detail

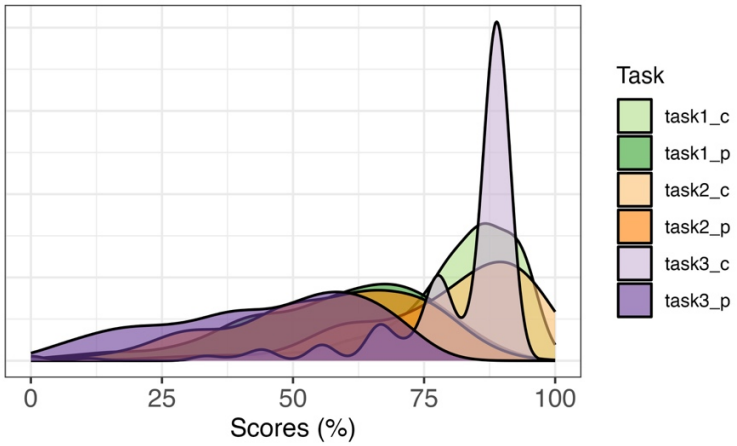
The first set of analyses were all completed using the manually derived scores and concerned whether points should be awarded for central and peripheral details or without regard to type of detail. First, I assessed whether we successfully replicated prior research showing that memory for central details is superior to memory for peripheral details. This would also suggest that memory for central and peripheral details differs, at least behaviorally.

Figure 1 displays score distributions for each of the story tasks by detail type. Descriptively, participants recalled a greater percentage of central details compared to peripheral details. Additionally, the distribution of performance for the central detail scores tended to be leptokurtic and negatively skewed; this is largely alleviated by summing recall performance without regard to type of detail (see Figure 2).

Given the distribution of performance, three Wilcoxon signed rank tests were conducted rather than *t*-tests. There was a significant difference in performance due to detail type in all three story recall tasks: story recall 1, $V = 36$, $p < .001$; story recall 2, $V = 14$, $p < .001$; and story recall 3, $V = 165$, $p < .001$. These results suggest that our conceptualization of central and peripheral details as well as our scoring is at least somewhat similar to other researchers' conceptualization and scoring of story recall (e.g., Sacripante et al., 2023).

Figure 1

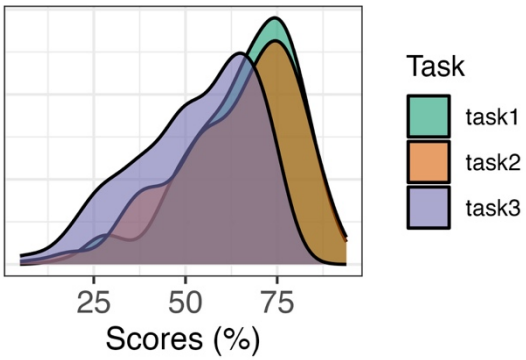
Distribution of Performance by Detail Type



Note. Detail type is denoted by the "_c" and "_p" subscripts, meaning central or peripheral detail, respectively.

Figure 2

Distribution of Performance by Task



Note. task1 = story task 1; task2 = story task 2; task3 = story task 3.

Confirmatory Factor Analyses

Next, I conducted confirmatory factor analyses to investigate whether and how memory for central and peripheral details differed. First, extreme outliers (greater than 3.5 standard deviations from the mean) were replaced with values equal to the cutoff—a total of 9 values (.21% of the data) were deemed outliers. Next, three participants were removed for being multivariate outliers (leaving 232 participants). There were some concerns about normality, thus I used maximum likelihood estimation with robust (Huber-White) standard errors. Missing values (n = 12) were addressed using full information likelihood.

Table 4 shows fit indices for several models. The first confirmatory analysis I conducted included factors estimating memory for peripheral details, memory for central details, fluid intelligence, working memory capacity, and crystallized intelligence. This model (subScore0 in Table 4) was non-positive definite. Next, I added residual correlations between subscores derived from the same tasks (i.e., memory for peripheral and central details from the same story), however, this model was also non-positive definite. Next, I explicitly set the correlation between the subscore factors equal to .99—this model was successfully estimated. The measurement model is shown in Figure 3; correlations amongst factors are shown in Figure 4.

Readers should note two things. First, the correlations amongst the factors are similar to correlations that have been reported in the literature examining individual differences in declarative memory (Martinez & Singleton, 2019; Unsworth, 2019; Wilhelm et al., 2013), providing some evidence of validity. Second, the correlations between the central and peripheral factors and the non-story factors are highly similar, suggesting that memory for central and peripheral details largely rely on the same cognitive processes. For example, the correlation between fluid intelligence and the peripheral memory factor is .69; with the central memory factor, the correlation is .67.

Given the fact that the correlation between the story recall factors is so large and the minor differences in correlations observed between these factors and other cognitive factors, I next estimated a model with a single story recall factor. The model was successfully estimated without any issues. Model fit was adequate (see "totalCFA" in Table 4). Comparing the AIC and BIC values between the subScore2 and totalCFA models, totalCFA is the more parsimonious of the two models.

Overall, the results of the latent variable analyses suggest that memory for central and peripheral details rely on similar cognitive processes. The correlation between the two factors is nearly one and correlations between these factors and other cognitive factors are nearly identical. Additionally, a model with a single story recall factor adequately fits the data, is more parsimonious, and also shows similar relationships with the other factors as both the peripheral and central memory factors.

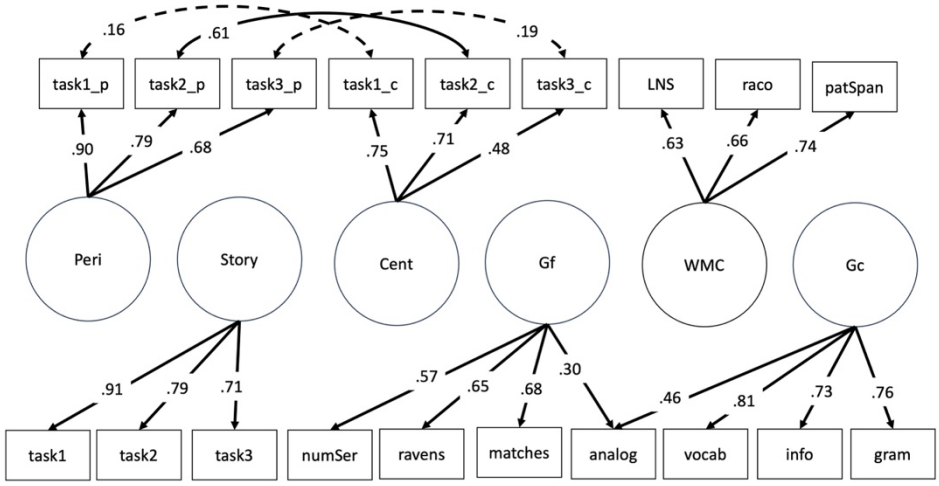
Table 4

Fit Statistics

	df	chisq	AIC	BIC	CFI	TLI	SRMR	RMSEA
subScore0	Non-positive definite							
subScore1	Non-positive definite							
subScore2	91	156.06	29181	29391	.96	.95	.05	.056 (.040, .070)
totalCFA	58	124.76	23878	24036	.95	.93	.05	.070 (.053, .087)

Figure 3

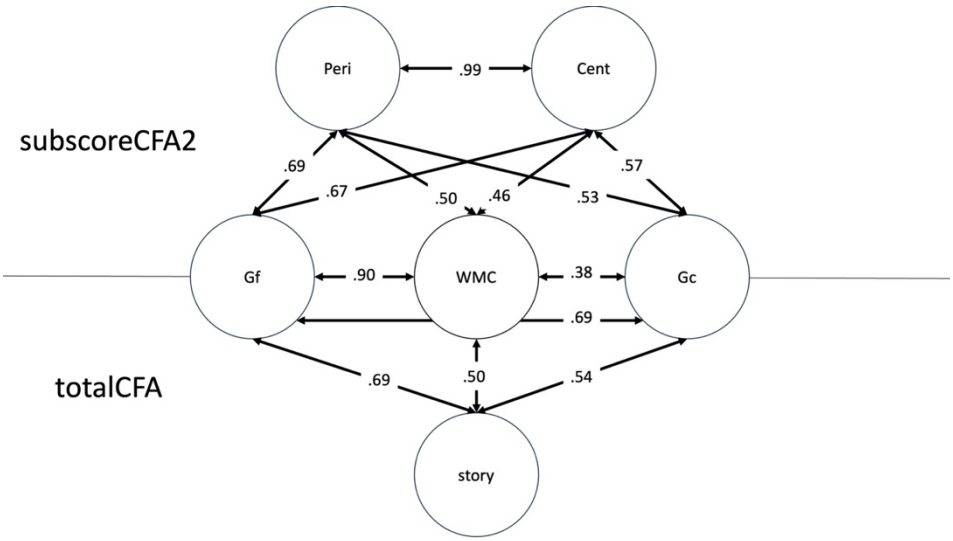
Task Loadings



Note. Solid paths are significant; dashed paths are not significant at the .05 level; Peri = factor representing memory for peripheral details; Cent = factor representing memory for central details; Story = story recall without regard for detail type; Gf = fluid intelligence; WMC = working memory capacity; Gc = crystallized intelligence; _c denotes central detail score; _p denotes peripheral detail score; LNS = letter-number sequence; raco = route and count; patSpan = pattern span; numSer = number series; analog = analogies; info = information; gram = grammar.

Figure 4

Confirmatory Factor Analyses



Note. SubscoreCFA2 and totalCFA models were estimated separately; here they are shown together to allow one to compare correlations across the models more easily.

Analysis and Results 2: Are automated approaches as good or better than manual scoring?

Confirmatory Factor Analyses

This set of analyses investigated whether story recall scoring could be automated using BERTScore (Zhang et al., 2020) and GPT-4 (OpenAI, 2023). Given the results of the analyses above with manual scoring, here, I disregarded detail type and only computed total scores. The same outlier removal and missing value treatment was used here as in the previous analyses.

The first analysis investigated correlations amongst factors estimated from BERTScore, GPT-4, and manually derived scores as well as with fluid intelligence, crystallized intelligence, and working memory capacity. This model also included residual correlations between the same task scored by the different models. The model was non-positive definite (Table 5, autoScore1). As the issue appeared to stem from the high correlation between the manual and GPT-4 derived factors, in the next model, the

correlation between these two factors was set to .99. This model (autoscore2) was also non-positive definite. Next, the model was simplified by removing other factors. This model, depicted in Figure 5, was successfully estimated and fit the data well (see Table 5, autoScore3); though note, the excellent fit is in large part due to the number of parameters estimated. Turning to the correlations amongst the story factors—are all very strong and, in the case of GPT-4 and manually derived scores, nearly perfect. Given the strength of these relations, one would expect that correlations with other variables will be similar across these memory factors, however, this is not assured (McCornack, 1956). To investigate correlations with the other factors, I estimated two confirmatory factor analyses, one for BERTScore and one for GPT-4.

The confirmatory factor analyses investigating scores derived from BERTScore and GPT-4 fit the data adequately (Table 5, BertScore and GPT-4). From these analyses, we can see that the correlations between the various scoring methods and the other factors are all similar (see Figure 6).

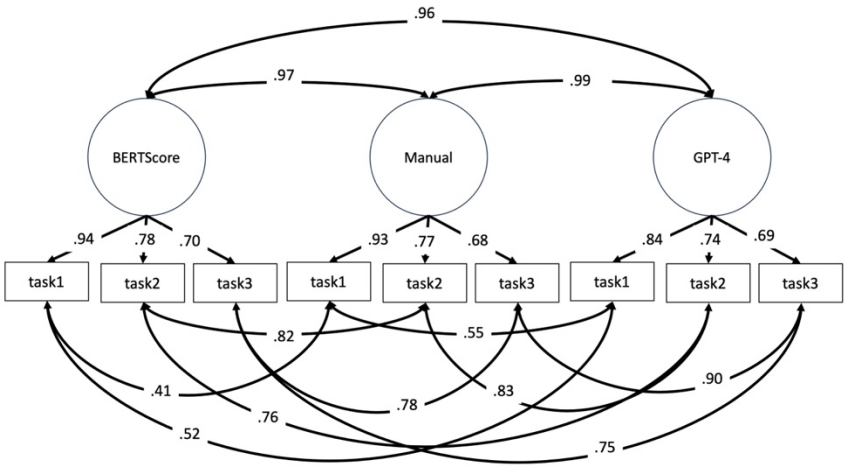
Table 5

Automated Scoring Models' Fit Statistics

	df	chisq	AIC	BIC	CFI	TLI	SRMR	RMSEA
autoScore1	Non-positive definite							
autoScore2	Non-positive definite							
autoScore3	16	15.24	13543	13674	1.0	1.0	.024	.000 (.000, .058)
BertScore	58	110.56	22595	22753	.96	.94	.045	.062 (.045, .080)
GPT-4	58	114.36	24078	24236	.95	.93	.046	.065 (.047, .082)

Figure 5

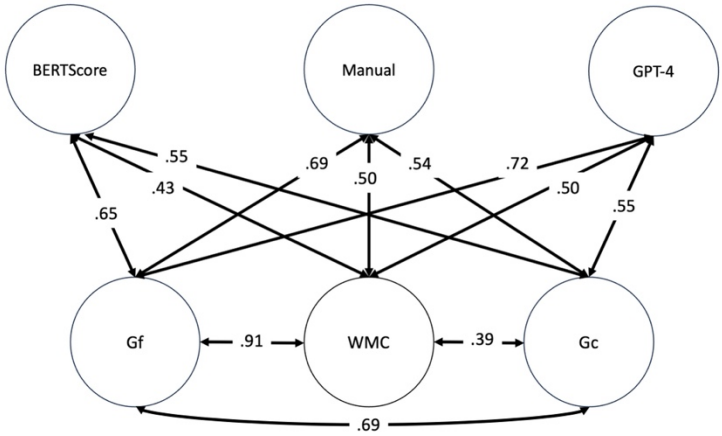
Confirmatory Factor Analysis



Note. task1, 2, and 3 under their respective factors are the same participant responses scored using the different methods.

Figure 6

Figure Depicting Relations Between Variously Scored Story Recall and Other Factors



Note. The story factors were estimated separately. All values shown are correlations. The "manual" factor is the same as "story" in Figure 4.

Further Investigations of the Adequacy of Automated Scoring Methods

The results of the confirmatory factor analyses suggest that one can obtain similar results by manually scoring story recall or by using BERTScore or GPT-4. However, in this study, each participant recalled nine stories and performance was aggregated into three task scores and then into a factor; differential psychologists may wish to use fewer stories. Thus, it is important to consider how the various scoring methods fared at the task and individual story level.

Table 6 displays descriptive statistics for each of the tasks and their component stories by scoring method as well as average correlations with the fluid intelligence, crystallized intelligence, and working memory capacity tasks.

From Table 6, we can observe that, depending on the scoring method, scores on some tasks and stories are highly leptokurtic. This positive kurtosis is generally resolved by averaging across stories to form task scores; however, this is not the case for the Story Task 1 set of stories that were scored by GPT-4. Turning to the correlations, though kurtosis can lead to spurious results, here, the correlations appear to be largely unaffected by violations of normality. For those interested, the entire correlation matrix is available in the Appendix as well as in the project repository at https://osf.io/5qxkh/.

Table 6

Descriptive Statistics and Correlations

	$\bar{x}$	*sd*	min	max	skew	kurt.	Gf	Gc	WMC
manual1	65.88	14.91	16.3	91.3	-0.83	0.39	0.40	0.37	0.30
bike_m	70.2	15.86	14.29	92.86	-0.77	0.59	0.32	0.31	0.27
early riser_m	63.32	17.24	0	88.24	-1.11	1.53	0.34	0.28	0.23
nurse_m	64.76	20.12	6.67	93.33	-0.74	0.07	0.32	0.32	0.24
BERTScore1	81.54	5.82	62.76	94.4	-0.74	0.83	0.36	0.38	0.26
bike_b	81.99	5.38	61.2	94.14	-0.59	1.21	0.34	0.32	0.26
early riser_b	83.37	7.56	42.43	97.24	-1.94	**7.3**	0.31	0.28	0.20
nurse_b	79.26	7.28	55.12	95.78	-0.74	1.09	0.35	0.37	0.23
GPT-4_1	83.73	13.27	23.91	95.65	-2.12	**4.86**	0.39	0.33	0.27
bike_g	86.78	14.21	14.29	100	-2.65	**7.99**	0.30	0.24	0.25
early riser_g	83.12	16.04	0	93.75	-2.81	**9.83**	0.33	0.28	0.20
nurse_g	87.12	19.53	3.33	100	-2.29	**4.91**	0.32	0.29	0.22
manual2	64.03	16.67	7.14	93.88	-0.74	0.05	0.36	0.38	0.26
balloon_m	64.92	20.99	0	100	-0.84	0.23	0.28	0.36	0.16
messenger_m	52.47	22.03	5.56	94.44	-0.37	-0.7	0.31	0.29	0.24
squeezer_m	77.26	18.84	0	100	-1.63	**3.34**	0.26	0.26	0.21
BERTScore2	78.19	6.29	50	90.81	-1.26	2.68	0.33	0.37	0.22
balloon_b	80.67	7.49	34.53	96.35	-1.69	**6.35**	0.27	0.31	0.14
messenger_b	74.03	6.95	43.84	89.24	-0.93	1.65	0.35	0.36	0.23
squeezer_b	80.22	7.38	35.24	94.47	-2.12	**7.77**	0.24	0.26	0.18
GPT-4_2	67.26	18.97	3.06	92.86	-1.12	0.75	0.33	0.34	0.24
balloon_g	78.15	26.55	0	100	-1.7	1.71	0.25	0.29	0.14
messenger_g	47.35	21.66	0	86.11	-0.35	-0.58	0.31	0.30	0.21
squeezer_g	79.87	23.87	0	100	-1.53	1.98	0.24	0.23	0.20
manual3	52.4	16.43	4.88	78.05	-0.6	-0.45	0.29	0.29	0.25
dog_m	53.36	19.76	0	78.57	-0.62	-0.49	0.23	0.25	0.18
fox_m	55.03	18.06	7.14	78.57	-0.69	-0.21	0.24	0.23	0.24
hen_m	48.53	18.78	7.69	76.92	-0.4	-0.75	0.29	0.30	0.25
BERTScore3	86.44	7.72	55.1	99.53	-0.83	0.94	0.29	0.30	0.23
dog_b	86.59	9.38	35	99.95	-1.8	**6.46**	0.24	0.23	0.18
fox_b	87.02	7.94	56.72	99.29	-0.77	0.42	0.28	0.28	0.23
hen_b	85.71	8.67	44.58	100	-1.12	2.19	0.27	0.27	0.20
GPT-4_3	62.32	22.55	2.44	100	-0.36	-0.56	0.30	0.32	0.23
dog_g	70.96	30.94	0	116.67	-0.15	-0.89	0.24	0.25	0.19
fox_g	77.01	28.9	0	116.67	-0.64	-0.48	0.25	0.24	0.18
hen_g	70.86	32.92	0	118.18	-0.02	-0.95	0.29	0.30	0.24

Note. Kurtosis values equal to or greater than 3 are bolded; min = minimum value observed in the data; max = maximum value observed in the data; kurt. = kurtosis; Gf = fluid intelligence; Gc = crystallized intelligence; WMC = working memory capacity.

Discussion

In this study, I investigated 1) whether story recall responses should be scored separately for recall of central and peripheral details and 2) whether responses could be scored automatically using computational methods. In this section, I discuss the results, their implications, note potential problems, and offer suggestions for future research.

Should Responses be Scored Separately for Recall of Central and Peripheral Details?

Based on the results of this study, it appears that memory for central and peripheral details largely rely on the same cognitive processes. Although participants recalled a statistically significant larger proportion of central details relative to peripheral details, the correlation between factors representing memory for the two types of details was nearly one. Additionally, correlations between these two factors and other factors were quite similar. The conclusion to be drawn from these results is that differential psychologists can *likely* disregard detail type, as little will be gained.

Of course, it is necessary to replicate and expand upon this research. One suggestion for future research examining individual differences in memory for central and peripheral details is to ask individuals to *either* recall central details or peripheral details without necessarily requiring that the details be written within a cohesive narrative. It may be that individuals do remember specific details, but they do not necessarily recall how they fit into a narrative. This may reduce the correlation between memory for central and peripheral details because recall of peripheral details would no longer depend on the recall of the central details which make up the plot.

A second suggestion would be to consider different approaches to scoring recall. In this study, recalled details were scored leniently: one point was awarded when a detail was recalled exactly as it appeared in the story or when the recalled detail was highly similar to the original detail; half points were awarded when recalled details were reasonably similar or, in the case of noun phrases, partially recalled. The rationale for using lenient scoring was twofold. First, to more closely replicate prior

research (Sacripante et al., 2023) and second, to maintain a contrast between story recall and other memory paradigms that traditionally require verbatim recall (e.g., paired-associate learning). However, it is possible that stricter scoring—particularly of the peripheral details which tend to be more concrete and specific—would result in a different pattern of correlations than that observed here.

Can Story Recall Scoring be Automated?

Manually scoring hundreds or potentially thousands of participant responses is time consuming and introduces subjectivity that may harm reliability and validity; fortunately, and in line with prior research (Chandler et al., 2019, 2021; Dunn et al., 2002), the results of this study suggest that one can use computational methods to automate scoring. Correlations amongst story recall factors derived from manual scoring as well as using GPT-4 and BERTScore were quite strong; moreover, these factors showed very similar correlations with fluid intelligence, crystallized intelligence, and working memory capacity. These results suggest that researchers can use automated approaches to obtain scores that are as valid as manually derived scores. However, researchers should be aware of the fact that the automated approaches used here did occasionally produce non-normally distributed scores.

It was hypothesized that reliability and validity could be improved by automation (see also Chandler et al., 2019, 2021), however, validity estimates (i.e., correlations) were all fairly similar across methods and, by implication, so was reliability. The increase in reliability and validity was to stem from the increased objectivity and precision that automated scoring offers. In the case of GPT-4 (and GPT 3.5), I was unable get scores more granular than half-points, thus presumably any increases in validity would have been a result of an increase in consistency in scoring. In the case of BERTScore, scores could differ by 1/10,000. Again, however, it appears that the automated approaches did no better than manual scoring, as correlations were quite similar.

It may be that there is little to gain by the increased consistency and precision offered by automated approaches. This may be because omissions are far more likely than semantic errors (Davis et

al., 2015; Jefferies et al., 2004). Both humans and language models will assign the lowest possible score to omitted details. Additionally, in the case of manual scoring, consistency will be high when scoring omitted details, as subjectivity does not factor in (though, of course, human error still does). Regarding semantic errors, it is likely that the types of semantic errors individuals commit are rather constrained when the material to be recalled is realistic and cohesive (Cofer, 1943). As an example, consider a story about a fox chasing a hen. It is possible that the animals will be substituted for other semantically related animals (e.g., a wolf is substituted for the fox) but as noted above, these errors will be relatively uncommon. It is also highly unlikely that other, more semantically distant animals will make appearances (e.g., a wolverine or python). If individuals make few and fairly similar semantic errors, then the impact of receiving .8348 or .5 points for a given detail will be fairly small, as the *ordering* of scores will be similar across the humans and language models. This is not to say that the increase in precision and consistency has *no* impact—rather, the impact may be too small to be detectable or of interest in basic research. Of course, researchers should further investigate this issue by, for example, assessing correlations with other declarative memory paradigms or by assessing how the scoring approaches fair in discriminating between groups of participants (e.g., as in clinical research)—situations in which small gains in precision may yield larger effects.

Though the automated approaches performed similarly to the manually derived scores in terms of correlations with other factors assessed here, BERTScore and GPT-4 occasionally generated leptokurtic distributions. This is a concern because severe kurtosis can lead to spurious results. Researchers interested in using an automated approach to score story recall may wish to take some steps to reduce the effects or likelihood of non-normality. The effects of non-normally distributed scores can be mitigated by employing large sample sizes and using robust statistics. It may also be possible to circumvent the issue of non-normality altogether. First, researchers may be able to use automated approaches with different parameters, prompts, or models. Researchers could potentially explore

different methods using the participant generated data from this study before attempting to generalize to a new data set. A second possibility for researchers interested in using automated approaches is to select the stories from this study that resulted in the most psychometrically sound scores. Regardless of the approach taken, researchers should be prepared to score at least a subset of responses manually to assess the performance of the automated approach of their choosing.

Given the possibility of non-normality, researchers interested in using story recall may wish to sacrifice efficiency and continue scoring responses manually. As noted previously, manual scoring does take time and it also introduces subjectivity, however, these issues may not be as significant a barrier as researchers think. Although a detailed record of the time it took to manually score responses was not kept, I estimate it took about one hour to train research assistants; about 10 hours for each research assistant to score the set of responses used to estimate reliability; and an additional 10 hours for each research assistant to score their set of responses. Thus, it took five individuals a combined total of about 105 hours to score a little over 2100 responses—in our case, this work was completed in about 3 weeks. Moreover, using lenient and therefore more subjective scoring, reliability was quite strong, with correlation coefficients amongst the raters always greater than .7. Still, the automated approaches were much more efficient. It took about five minutes to score all responses using BERTScore and about one hour using GPT-4 (though note, scoring was done in October 2023 when OpenAI was limiting the rate of requests to GPT-4 to 200 per minute). Researchers may also be interested in knowing that BERTScore was free to use while the cost of scoring the responses with GPT-4 was about $14.

Summary and Conclusion

To summarize, there are two major conclusions to be drawn from this study. First, memory for peripheral and central details appears to rely on the same cognitive processes. While participants recalled a larger proportion of central details compared to peripheral details, the correlation between the two memory factors was remarkably strong. Additionally, their relationships with other cognitive

factors exhibited considerable similarity. Consequently, differential psychologists interested in story recall can disregard detail types, as little additional insight may be gained from treating them separately. Nonetheless, it is crucial to replicate and expand on this research, potentially by asking participants to list recalled details rather than writing them in a cohesive narrative.

Secondly, with respect to the potential automation of story recall scoring, this study revealed that computational methods, such as GPT-4 and BERTScore, can be employed to streamline the scoring process. The correlations among manual scoring and automated methods were strong, with no significant discrepancy observed in their relationships with fluid intelligence, crystallized intelligence, and working memory capacity. Though researchers should note that the automated approaches did occasionally produce non-normally distributed scores. Researchers interested in automated scoring should consider steps to mitigate the effects of deviations from normality or explore other parameters, prompts, or models to reduce the likelihood of generating kurtotic score distributions in the first place.

Declarations

Funding

Funding for this work was provided by an Internal Research and Development grant from ARLIS as well as a grant from the Office of Naval Research (N00014-21-1-2220).

Conflicts of Interest

None.

Ethics approval

This study was approved by the University of Maryland, College Park Internal Review Board.

Consent to participate

All participants (re-)consented to participate at the beginning of their session.

Consent for publication

N/A

Availability of data and materials

All data and materials are available here: https://osf.io/5qxkh/

Code availability

R scripts used to interact with BERTScore and GPT-4 as well as analyses are available here:

https://osf.io/5qxkh/

References

Abikoff, H., Alvir, J., Hong, G., Sukoff, R., Orazio, J., Solomon, S., & Saravay, S. (1987). Logical memory subtest of the wechsler memory scale: Age and education norms and alternate-form reliability of two scoring systems. *Journal of Clinical and Experimental Neuropsychology*, *9*(4), 435–448. https://doi.org/10.1080/01688638708405063

Allen, R. J., & Baddeley, A. D. (2009). Working memory and sentence recall. In *Interactions between short-term and long-term memory in the verbal domain* (pp. 63–85). Psychology Press. https://books.google.com/books?hl=en&lr=&id=Mhx5AgAAQBAJ&oi=fnd&pg=PA63&dq=Working+memory+and+sentence+recall&ots=bfgMGOMkSd&sig=im6smgrBoZSHV3MaIwpQ2CidJl8

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language*, *61*(3), 438–456.

Baddeley, A. D., & Wilson, B. A. (2002). Prose recall and amnesia: Implications for the structure of working memory. *Neuropsychologia*, *40*(10), 1737–1743.

Baek, M. J., Kim, H. J., Ryu, H. J., Lee, S. H., Han, S. H., Na, H. R., Chang, Y., Chey, J. Y., & Kim, S. (2011). The usefulness of the story recall test in patients with mild cognitive impairment and Alzheimer's

disease. *Aging, Neuropsychology, and Cognition*, *18*(2), 214–229. https://doi.org/10.1080/13825585.2010.530221

Bartsch, L. M., & Oberauer, K. (2021). The effects of elaboration on working memory and long-term memory across age. *Journal of Memory and Language*, *118*, 104215.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238. https://doi.org/10.1016/0160-2896(84)90009-6

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, *10*(1), 3–47. https://doi.org/10.1016/0273-2297(90)90003-M

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behavior*, *28*, 41–100.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bunting, M., Cowan, N., & Scott Saults, J. (2006). How does running memory span work? *Quarterly Journal of Experimental Psychology*, *59*(10), 1691–1700. https://doi.org/10.1080/17470210600848402

Cassady, J., & Chittooran, M. M. (2010). Wechsler Memory Scale–Fourth Edition. *WMS-IV*. Mental Measurements Yearbook with Tests in Print. https://search.ebscohost.com/login.aspx?direct=true&db=mmt&AN=test.3181&site=ehost-live

Chandler, C., Foltz, P., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., Cohen, A. S., Holmlund, T. B., & Elvevåg, B. (2019). Overcoming the bottleneck in traditional assessments of verbal memory: Modeling

human ratings and classifying clinical group membership. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 137–147. https://aclanthology.org/W19-3016/

Chandler, C., Holmlund, T. B., Foltz, P. W., Cohen, A. S., & Elvevåg, B. (2021). Extending the usefulness of the verbal memory test: The promise of machine learning. *Psychiatry Research*, *297*, 113743.

Cofer, C. N. (1943). An analysis of errors made in the learning of prose materials. *Journal of Experimental Psychology*, *32*(5), 399.

Cunje, A., Molloy, D. W., Standish, T. I., & Lewis, D. L. (2007). Alternate forms of logical memory and verbal fluency tasks for repeated testing in early cognitive changes. *International Psychogeriatrics*, *19*(1), 65–75.

Davis, D. K., Alea, N., & Bluck, S. (2015). The difference between right and wrong: Accuracy of older and younger adults' story recall. *International Journal of Environmental Research and Public Health*, *12*(9), 10861–10885.

Della Sala, S., Gray, C., Baddeley, A., Allamano, N., & Wilson, L. (1999). Pattern span: A tool for unwelding visuo–spatial memory. *Neuropsychologia*, *37*(10), 1189–1199.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A., & Flicker, L. (2002). Latent Semantic Analysis: A New Method to Measure Prose Recall. *Journal of Clinical and Experimental Neuropsychology*, *24*(1), 26–35. https://doi.org/10.1076/jcen.24.1.26.965

Ekstrom et al. (1976). *Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.*

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309.

Epskamp, S. (2015). semPlot: Unified Visualizations of Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 474–483. https://doi.org/10.1080/10705511.2014.937847

Foldi, N. S. (2011). Getting the hang of it: Preferential gist over verbatim story recall and the roles of attentional capacity and the episodic buffer in Alzheimer disease. *Journal of the International Neuropsychological Society*, *17*(1), 69–79.

Gerver, C. R., Griffin, J. W., Dennis, N. A., & Beaty, R. E. (2023). Memory and creativity: A meta-analytic examination of the relationship between memory systems and creative cognition. *Psychonomic Bulletin & Review*, 1–39.

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, *26*(1), 69–83.

Gonthier, C., & Thomassin, N. (2015). Strategy use fully mediates the relationship between working memory capacity and performance on Raven's matrices. *Journal of Experimental Psychology: General*, *144*(5), 916–924. https://doi.org/10.1037/xge0000101

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* (arXiv:2006.03654). arXiv. http://arxiv.org/abs/2006.03654

Holmlund, T. B., Chandler, C., Foltz, P. W., Cohen, A. S., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., & Elvevåg, B. (2020). Applying speech technologies to assess verbal memory in patients with serious mental illness. *NPJ Digital Medicine*, *3*(1), 33.

Hughes, M. M., Karuzis, V. P., Kim, S., O'Rourke, P., Sumer, A., Liter, A., Bunting, M. F., & Campbell, S. G. (2016). *Assessing aptitude for USAF cyber warfare operations training: Interim results from field testing*. University of Maryland Center for Advanced Study of Language.

Jefferies, E., Ralph, M. A. L., & Baddeley, A. D. (2004). Automatic and controlled processing in sentence recall: The role of long-term and working memory. *Journal of Memory and Language*, *51*(4), 623–643.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.

Kessels, R. P., Van Zandvoort, M. J., Postma, A., Kappelle, L. J., & De Haan, E. H. (2000). The Corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, *7*(4), 252–258.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). *Constraint Relaxation and Chunk Decomposition in Insight Problem Solving*. 22.

Mandler, J. M., Scribner, S., Cole, M., & DeForest, M. (1980). Cross-cultural invariance in story recall. *Child Development*, 19–26.

Martinez, D., & Singleton, J. L. (2018). Predicting sign learning in hearing adults: The role of perceptual-motor (and phonological?) processes. *Applied Psycholinguistics*, *39*(5), 905–931.

Martinez, D., & Singleton, J. L. (2019). The effect of bilingualism on lexical learning and memory across two language modalities: Some evidence for a domain-specific, but not general, advantage. *Journal of Cognitive Psychology*, *31*(5–6), 559–581. https://doi.org/10.1080/20445911.2019.1634080

McCornack, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology*, *40*(5), 343.

Mielicki, M. K., Koppel, R. H., Valencia, G., & Wiley, J. (2018). Measuring working memory capacity with the letter–number sequencing task: Advantages of visual administration. *Applied Cognitive Psychology*, *32*(6), 805–814.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*.

Ooms, J. (2014). *The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects* (arXiv:1403.2805). arXiv. http://arxiv.org/abs/1403.2805

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193.

Power, D. G., Logue, P. E., McCarty, S. M., Rosenstiel, A. K., & Ziesat, H. A. (1979). Inter-rater reliability of the russell revision of the wechsler memory scale: An attempt to clarify some ambiguities in scoring. *Journal of Clinical Neuropsychology*, *1*(4), 343–345. https://doi.org/10.1080/01688637908401109

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Rabin, L. A., Paré, N., Saykin, A. J., Brown, M. J., Wishart, H. A., Flashman, L. A., & Santulli, R. B. (2009). Differential Memory Test Sensitivity for Diagnosing Amnestic Mild Cognitive Impairment and Predicting Conversion to Alzheimer's Disease. *Aging, Neuropsychology, and Cognition*, *16*(3), 357–376. https://doi.org/10.1080/13825580902825220

Ratner, H. H., Schell, D. A., Crimmins, A., Mittelman, D., & Baldinelli, L. (1987). Changes in adults' prose recall: Aging or cognitive demands? *Developmental Psychology*, *23*(4), 521.

Raven et al. (1998). *Raven, J. C., & John Hugh Court. (1998). Raven's progressive matrices and vocabulary scales (Vol. 759). Oxford: Oxford pyschologists Press.*

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research*. https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research

Robinson-Whelen, S., & Storandt, M. (1992). Immediate and delayed prose recall among normal and demented adults. *Archives of Neurology*, *49*(1), 32–34.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432.

Ryan, J. J., & Gontkovsky, S. T. (2023). Age effects on Wechsler Memory Scale–Fourth Edition (WMS–IV) indexes and subtests. *Applied Neuropsychology: Adult*, 1–7. https://doi.org/10.1080/23279095.2023.2217461

Sacripante, R., Logie, R. H., Baddeley, A., & Della Sala, S. (2023). Forgetting rates of gist and peripheral episodic details in prose recall. *Memory & Cognition*, *51*(1), 71–86. https://doi.org/10.3758/s13421-022-01310-5

Schnabel, R. (2012). Overcoming the Challenge of Re-assessing Logical Memory. *The Clinical Neuropsychologist*, *26*(1), 102–115. https://doi.org/10.1080/13854046.2011.640642

Strong, E., DiGiammarino, A., Weng, Y., Kumar, A., Hosamani, P., Hom, J., & Chen, J. H. (2023). Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Internal Medicine*, *183*(9), 1028–1030.

Thurstone, L. L. (1938). *Primary mental abilities* (Vol. 119). University of Chicago Press Chicago.

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79. https://doi.org/10.1037/bul0000176

Ushey, K., Allaire, J. J., & Tang, Y. (2020). reticulate: Interface to'Python'. *R Package Version*, *1*, 18.

Wechsler, D. (1997). *Wechsler Memory Scale—Third Edition (WMS–III) technical and interpretive manual*. Psychological Corporation.

Wechsler, D. (2009). *Wechsler Memory Scale—Fourth Edition (WMS–IV) technical and interpretive manual*. Pearson.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*, 1–29.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wickham, H. (2020). httr: Tools for Working with URLs and HTTP. *R Package Version*, *1*(2).

Wickham, H., François, R., Henry, L., & Müller, K. (2015). dplyr: A grammar of data manipulation. *R Package Version 0.4*, *3*, p156.

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 433.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT* (arXiv:1904.09675). arXiv. http://arxiv.org/abs/1904.09675

APPENDIX

Task Descriptives

task	n	mean	sd	min	max	skew	kurtosis	IC[a]
story1_c	235	81.72	11.58	33.33	93.33	-1.42	2.25	.62
story1_p	235	58.22	17.79	4.84	90.32	-0.67	0.08	.86
story2_c	234	79.2	16.58	9.38	100	-1.19	1.55	.75
story2_p	234	56.68	17.8	6.06	90.91	-0.46	-0.48	.86
story3_c	234	81.23	14.54	0	88.89	-2.89	10.3	.75
story3_p	234	44.29	18.98	3.12	75	-0.41	-0.92	.90
story1	235	65.88	14.91	16.3	91.3	-0.83	0.39	.88
story2	234	64.03	16.67	7.14	93.88	-0.74	0.05	.90
story3	234	52.4	16.43	4.88	78.05	-0.6	-0.45	.91
bert1	235	81.54	5.82	62.76	94.4	-0.74	0.83	---[b]
bert2	234	78.19	6.29	50	90.81	-1.26	2.68	---[b]
bert3	234	86.44	7.72	55.1	99.53	-0.83	0.94	---[b]
gpt1	235	83.73	13.27	23.91	95.65	-2.12	4.86	---[b]
gpt2	234	67.26	18.97	3.06	92.86	-1.12	0.75	---[b]
gpt3	234	62.32	22.55	2.44	100	-0.36	-0.56	---[b]
analogies	235	73.12	14.23	27.78	100	-0.49	-0.37	.64
numSeries	235	67.6	18.34	13.33	100	-0.27	-0.64	.75
ravens	235	58.7	16.67	11.11	100	-0.15	-0.26	.72
matches	234	47.57	10.69	0	76	-0.8	2.19	.61
vocab	235	61.05	12.74	27.08	91.67	-0.12	-0.39	.79
gram	235	48.33	20.86	4.76	95.24	0.2	-0.68	.8
info	235	69.21	13.85	25	97.5	-0.72	0.32	.82
lns	231	60.97	12.7	20.91	93.06	-0.14	0.07	.82
raco	233	60.96	19	15.83	98.33	-0.09	-0.9	.75
patSpan	234	62.91	16.91	0	94.44	-0.41	0.03	.76

Note. [a]IC = internal consistency, estimated as Cronbach's Alpha; [b]reliability was not estimated.

Pearson Correlations

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	story1_c																								
2	story1_p	.71																							
3	story2_c	.49	.60																						
4	story2_p	.58	.69	.81																					
5	story3_c	.39	.41	.34	.34																				
6	story3_p	.53	.58	.44	.50	.42																			
7	story1	.83	.98	.61	.71	.43	.60																		
8	story2	.58	.69	.91	.98	.36	.50	.71																	
9	story3	.55	.60	.46	.51	.58	.98	.62	.52																
10	bert1	.76	.88	.61	.69	.44	.58	.90	.70	.61															
11	bert2	.56	.65	.82	.86	.32	.49	.67	.89	.50	.69														
12	bert3	.52	.62	.48	.53	.58	.84	.63	.54	.87	.65	.54													
13	gpt1	.80	.82	.56	.64	.41	.51	.86	.64	.54	.85	.60	.54												
14	gpt2	.57	.65	.86	.88	.36	.47	.67	.91	.49	.66	.86	.52	.63											
15	gpt3	.54	.61	.46	.51	.63	.91	.63	.52	.94	.62	.49	.86	.54	.49										
16	analogies	.33	.39	.41	.43	.15	.25	.40	.44	.26	.38	.39	.25	.39	.39	.27									
17	numSeries	.22	.31	.18	.25	.16	.18	.30	.24	.20	.34	.26	.26	.30	.23	.21	.40								
18	ravens	.37	.43	.29	.40	.23	.33	.44	.38	.34	.41	.34	.33	.43	.37	.36	.37	.36							
19	matches	.37	.40	.33	.35	.19	.35	.42	.36	.36	.40	.33	.32	.44	.34	.36	.38	.37	.48						
20	vocab	.28	.35	.38	.33	.14	.31	.35	.36	.31	.37	.34	.30	.31	.36	.31	.50	.23	.31	.32					
21	gram	.35	.40	.33	.34	.19	.27	.41	.35	.28	.42	.38	.28	.37	.33	.29	.49	.43	.39	.30	.64				
22	info	.21	.28	.36	.33	.18	.33	.28	.36	.33	.31	.34	.34	.27	.33	.35	.52	.32	.35	.44	.57	.47			
23	lns	.28	.33	.26	.31	.13	.21	.33	.31	.21	.34	.29	.23	.29	.28	.21	.38	.49	.39	.34	.24	.37	.18		
24	raco	.20	.27	.13	.24	.13	.26	.27	.22	.26	.19	.17	.20	.23	.18	.23	.25	.30	.33	.38	.11	.15	.13	.42	
25	patSpan	.26	.26	.22	.25	.16	.29	.27	.25	.29	.24	.21	.26	.29	.25	.26	.29	.35	.40	.49	.17	.20	.27	.41	.50

Note. bert1, gpt1, and others refer to Story 1 as scored by BERTScore and GPT-4, respectively.