Distinct paths to false memory revealed in hundreds of narrative recalls

Phoebe Chen[1,†,*], **Omri Raccah**[2,†*], **Vy A. Vo**[3, 4, ‡], **David Poeppel**[1,5], **and Todd M. Gureckis**[1]

[1]Department of Psychology, New York University, New York, NY, USA
[2]Department of Psychology, Yale University, New Haven, CT, USA
[3]Intel Labs, Intel Corporation, Hillsboro, OR, USA
[4]Max Planck Institute for Software Systems, Saarbrücken, Germany
[5]Center for Language, Music, and Emotion, NYU (CLaME), New York, NY, USA
[†]denotes equal contribution
[‡]Now at Thomson Reuters Labs. This work was not performed as part of this position.
[*]corresponding authors: Phoebe Chen (hc2896@nyu.edu) and Omri Raccah (omri.raccah@yale.edu)

ABSTRACT

Memory distortions emerge from a complex interplay between prior knowledge and ongoing experience—dynamics which are not readily provoked in controlled laboratory experiments. Here we investigate naturally occurring memory distortions using the largest known dataset of narrative recall, comprising hundreds of spoken recollections. Using large language models (LLMs), we developed an automated pipeline to detect and classify spontaneous false memories. Across two validation experiments, we demonstrate that human-AI agreement matches inter-human reliability in detecting and cataloging memory distortions. We show that false memories reflect two distinct phenomena which are driven by separable semantic factors: similarity to prototypical narrative patterns drives factual errors (distortions of actual content), whereas contextual surprise drives confabulations (entirely fabricated details). Through this combination of large-scale naturalistic data and AI-powered automation tools, we reveal memory processes that controlled laboratory paradigms cannot easily capture and illuminate the complex dynamics of human (mis)remembering in real-world contexts.

SIGNIFICANCE

The way that our memories distort the truth influences many aspects of society, such as the veracity of eyewitness testimony or our susceptibility to misinformation. The present paper asks if there are predictable aspects of our experience that influence these distortions. False memory has largely been studied using highly controlled list-learning paradigms. While these paradigms have been invaluable, they offer limited insight into how distortions arise during natural recollection. Here, we investigate spontaneously occurring false memories in the spoken recall of narratives. We introduce an automated approach using LLMs to detect and classify memory errors across hundreds of participants. Our validated findings illustrate how false memories are not a unitary phenomenon but comprise distinct error types: similarity to prototypical narrative patterns predicts factual errors (distortions of actual content), while contextual surprise drives confabulations (entirely fabricated details). This dissociation provides direct empirical support for theories proposing distinct cognitive mechanisms of memory reconstruction. Overall, our work demonstrates that naturalistic paradigms can uncover memory processes difficult to study using small-scale laboratory studies.

Introduction

Far from being veridical records of our lives, our memories are distorted reflections of the truth. In his seminal study of narrative recall, Bartlett[1] showed that people distorted story details over repeated retellings—"canoes" became "boats," "five men" became "some warriors." They also generated fictitious details consistent with prior events or knowledge, such as replacing "something black" with "blood" or fabricating that the man died "at sunset." These errors illustrate that memory is an active process of reconstructing encoded information, rather than a passive repository of details.

Studying distortions and errors in memory, especially those that occur under naturalistic conditions, is exceedingly difficult. One reason is that memory errors can be rare and highly variable across individuals. As a result, small-sample laboratory studies might fail to observe substantial variance in memory errors, and lack the ability to detect the situational patterns that cause them. While certain forms of false memory are reliably elicited in laboratory settings[2,3], extending these findings to more naturalistic settings presents additional challenges. Naturalistic memory studies typically employ small samples (20–30 participants) to enable manual scoring of the recall by human raters, and lack statistical power to detect infrequent memory distortions. In fact, it can be difficult even to reliably define what constitutes an memory error in everyday life because memory of experiences are rarely retrieved exactly as they occurred. Therefore, addressing false memories under naturalistic conditions demands large-scale datasets and novel methodological approaches to annotate and score the data.

In the present study we leverage two new methodological innovations to address our primary goal of understanding memory errors in naturalistic recall. First, we analyze a recently published large-scale dataset consisting of hundreds of spoken narrative recollections[4]. This dataset contains professionally coded transcripts of hundreds of individuals retelling a narrative story from memory. Analyzing such a large, unstructured data set would be expensive and time-consuming. However, our second innovation is that we leverage state-of-the-art large language models (LLMs), which are trained using in-context learning to detect and classify memory errors. We validate this approach showing that human-AI agreement is at least a high as inter-human agreement.

The current work builds upon a century of psychological research that has used controlled experimental paradigms to explore how these memory distortions occur, and what kinds of errors can be observed. Forgetting encompasses errors of omission, while other errors commonly described as false memories are errors of commission[5,6]. Commission errors have been further split into subtypes[7], mainly by considering how they may arise from different steps of the reconstructive process of memory recall. For example, binding the contents of memory to the wrong context results in misattribution errors[8,9]. This includes errors that attribute imagined content (e.g., fabricated childhood events) to the real past, a phenomenon robustly studied by a set of paradigms iterating on the notion of 'imagination inflation' or 'misinformation'[10–13]. Another type of error might involve a memory that has been weakly encoded, or encoded along with a strong overarching structure—these memories may be recalled with newly introduced details that are consistent with the person's beliefs or the overarching structure of the memory, but are unfaithful to the original inputs[14]. The classic Deese-Roediger-McDermott (DRM) paradigm demonstrates this principle: participants study lists of semantically related words (e.g., bed, rest, dream) and subsequently show robust false recognition for non-presented but related lures (e.g., sleep), often at rates comparable to actually presented items[2,15,16]. These effects extend beyond word lists to semantically rich images (e.g., farm scenes) and occur across both recognition and free-recall paradigms[2]. While these controlled paradigms have provided crucial empirical tractability, it is unclear how and when they might generalize to real world situations[17–19].

Highly controlled paradigms oversimplify the complex, interacting semantic factors that drive memory reconstruction in everyday contexts. While recent studies have begun to address how accurate recall occurs in naturalistic settings[4,20–22], they have not addressed inaccurate, or distorted, memory. Methodological barriers have impeded this progress: controlled paradigms to study false memory rely on some experimental manipulation to encourage, or induce, a higher rate of memory distortions to provide enough statistical power to characterize the effect. In list learning studies, the experimental manipulation is the choice of structured stimulus sets, such as words that all belong to the same category. In misinformation studies, the manipulation is the introduction of inaccurate information about an already-encoded event. By contrast, naturalistic memory studies simply present information and prompt the subjects to recall the information in a realistic way, such as by freely recalling the events of a movie[20,23]. To move beyond these limitations, we apply recent advances in natural language processing to analyze naturalistic memory recalls. These tools can enable automated text stimulus annotation and other tasks in developing and analyzing psychological experiments[24]. For example, some work has employed large language models (LLMs) to generate narrative stimuli[25], predict human reading times[26], automate segmentation of narratives[27,28] and score recalls[4,25].

In this work, we investigate spontaneously occurring false memories in the naturalistic recollection of spoken narratives. We leverage our recently published dataset containing recollections from hundreds of participants across four spoken narratives[4]. The dataset has already been validated for the study of memory, as it has been used to replicate key findings in the broader memory literature[29]. Our work makes two main contributions with these data. First, we developed an automatic LLM-based pipeline to detect and label two types of memory distortions, factual errors and confabulations, in naturalistic story recollections. Across two experiments, we validate this approach through showing that these annotations are comparable to human raters.

Second, we test whether different semantic factors predict these memory distortions in recall, and rely on advances in models of natural language processing (NLP) to help define these metrics.

This work addresses a fundamental gap in our understanding of memory by examining false memories as they naturally occur during narrative recollection—an understudied yet ubiquitous aspect of human behavior. Our approach provides novel methodological tools for investigating naturally occurring memory distortions and reveals complex semantic dimensions that capture individual tendencies to misremember beyond traditional item-list paradigms.

Results

Narrative free recall paradigm

To investigate the spontaneous occurrence of memory distortions under naturalistic conditions, we sought a paradigm that (1) closely mimics real-world remembering—the retelling of narrative content and (2) provides enough data to reliably observe memory errors. Our recently published "Naturalistic Free Recall" Dataset[4] (NFRD) meets these demands: hundreds of online participants listened to a spoken narrative and immediately recalled it in as much detail as possible. To the best of our knowledge, this dataset represents the largest free recall dataset for naturalistic materials (see also[25]), making it particularly well-suited for the identification of naturally occurring false memories.

The stimuli comprised four distinct spoken narratives in English. Three of the narratives were sourced from The Moth Radio Hour podcast series, specifically Pieman, Eyespy, and Oregontrail. The fourth narrative was the first chapter from an audiobook, henceforth referred to as Baseball. On average, each audio clip lasted 11 minutes and 35 seconds (see Methods and Materials section). Each participant was presented with two of the four narrative stimuli included in the NFRD, and performed a spoken recall for at least 4 minutes following each story (Figure 1A). The dataset includes professionally curated, high-fidelity, transcripts of each narrative and all verbal recollections. We conducted analyses across the entire dataset, collapsing across the four narratives to increase statistical power (N = 229; 145 female; $Mean_{age}$ = 25.03, SD_{age} = 11.15). An detailed description of the dataset can be found in our published paper[4].

Annotating memory distortions with LLMs

In recent years, natural language processing (NLP) and LLMs have become powerful tools in studying narrative memory[24,30]. Specifically, generative LLMs have enabled automated segmentation of narratives[27,28] and scoring of recall quality[25]. We developed an automated pipeline using GPT-4o[31] to identify and categorize false memories in narrative recollections–a task that requires the model to distinguish between accurate recall and false memories based on comparison with source material. We relied on in-context learning to leverage the ability of modern LLMs to perform complex classification tasks when provided with detailed instructions and examples—avoiding the need for task-specific training or fine-tuning via backpropagation. This required the development and tuning of specific prompts to optimize LLM performance on this task (see Methods and Materials).

We first segmented original narratives into events[4,21], representing semantically coherent portions of the story. We also segmented the recalls into sentences to allow the LLM to annotate false memory details at a finer grain. This relied on a two-stage classification pipeline. First, an event-matching task identified which event of the original narrative each recall sentence referenced, or marked it as unmatched if no clear correspondence existed (1.56% of all sentences; see Methods & Materials). Second, a memory classification task evaluated each sentence for three types of content: factual errors (details contradicting the story), confabulations (fabricated information absent from the story), and inferences (plausible details derived from world knowledge). Both tasks contained the complete narrative text and detailed task instructions as context (Table S2, S3 and S4), allowing the LLM to make informed comparisons between participants' recollections and the source material.

Sentences containing factual errors or confabulations were classified as false memories, while inferences were treated as an instance of true recollection. We included inferences for two reasons. First, prior research shows that inferences often reflect successful reconstructive retrieval[32], where new information is combined with past experiences. By including an inference category, we aim to separate adaptive reconstruction from memory errors[33,34]. Second, this approach follows a common strategy in machine learning: introducing an "unknown" or a "catch" category to handle ambiguous cases[35,36]. In practice we found that it improved the classification of factual errors and confabulations.

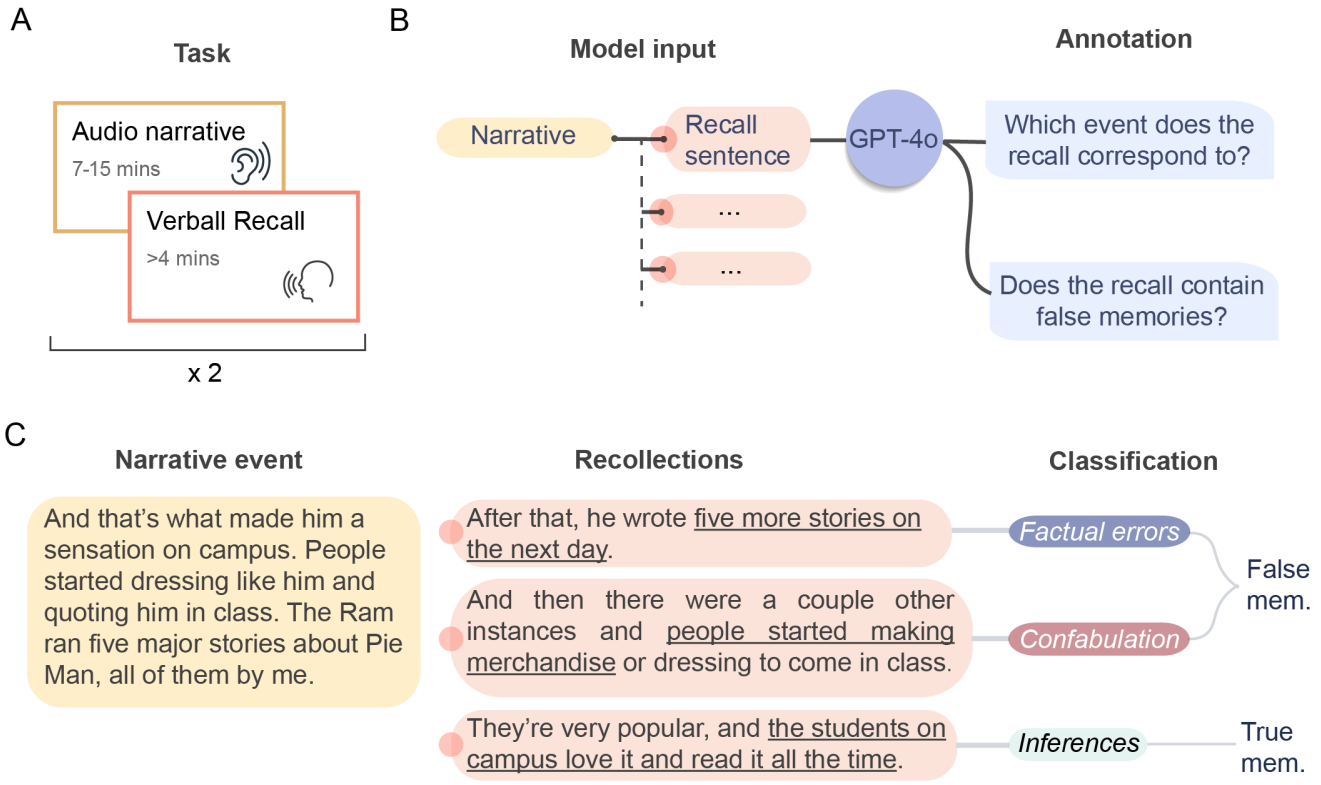

Figure 1. Task design and false memory annotation pipeline. **(A)** Task design of the narrative free recall paradigm[4]. Participants listened to two spoken stories and immediately recalled them verbally for at least four minutes; we then performed automatic speech-to-text transcription with professional review of the audio files. **(B)** Recall annotation approach. We employed in-context learning with GPT-4o to perform two consecutive tasks: (1) matching recall sentences to corresponding story events, and (2) classifying types of memory. We defined three distinct memory types based on prior theoretical frameworks[7,37,38]: factual errors (details that contradict the original narrative), confabulations (fabricated information not present in the story), and inferences (plausible details derived from world knowledge or reasonable interpretation). Full prompts for both tasks are attached in Table S2, S3 and S4. **(C)** Example classification outputs. An example story event from Pieman (left) is shown alongside three sample recollections from different participants (center) that were matched to this event. The LLM was instructed to return the specific portions of the recollections classified as factual errors (blue), confabulations (red), or true-memory inferences (light green).

Human validation

Next we sought to validate the performance of our annotation pipeline. Validating LLM-based classifications of memory phenomena presents unique challenges, as participants' recollections lack an objective ground-truth and their relation to the narratives rely on their personal interpretations. These challenges have sparked important discussions around how to validate the performance of LLMs when identifying and labeling complex psychological phenomena. Demszky et al.[39] argue that validation of LLMs against human behaviors should account for inherent human variability, rather than assume a single ground truth. In line with this perspective, recent work evaluates LLMs by comparing human-AI agreement to inter-rater agreement, effectively treating the LLM as a potential proxy for average human judgments[25,27]. This approach reframes human-AI validation as a question of alignment with collective human interpretation, rather than with a single ground truth label. Importantly, recent work has applied this approach to distinct aspects of narrative memory, including autobiographical details[40], event boundary detection[27] and empathy judgments[41].

To apply this approach to study of spontaneously occurring false memories, we conducted two validation experiments comparing human-human agreement (inter-rater reliability) with human-AI agreement. In the following experiments, two cohorts of participants evaluated the scoring of recall sentences after listening to the original narratives.

Validation Experiment 1: Open-ended classification

Twenty-seven participants evaluated 60 recall sentences (30 per story) in an open-ended format. Each sentence was presented with the corresponding (matched) story context and the AI's binary classification (accurate vs. inaccurate), with incorrect segments underlined in false memory trials. Participants judged whether they agreed with the classification and reported their confidence levels. This validation did not distinguish between factual errors and confabulations (see Experiment 2). To avoid bias, we only told participants that they were evaluating another rater, and did not specify that it was an LLM.

We calculated both inter-subject agreement (the proportion of trials on which pairs of human raters made identical judgments) and subject–AI agreement (the proportion of trials on which each human rater's judgment matched the AI's classification). We used a nonparametric bootstrapping procedure to perform statistical comparisons and error estimates (see section Human validation in Methods for details). We found that subject-AI agreement significantly exceeded inter-subject agreement for both accurate trials (subject-AI: 0.71 [95% CI: 0.66, 0.76] vs. inter-subject: 0.62 [95% CI: 0.57, 0.66], $p < 0.001$) and false memory trials (subject-AI: 0.66 [95% CI: 0.60, 0.72] vs. inter-subject: 0.60 [95% CI: 0.57, 0.64], $p = 0.002$, Figure 2A). This suggests that the AI's ratings more closely approximated the average human judgment than any individual participant.

Validation Experiment 2: Comprehension-based validation

The findings from the first experiment show that AI ratings were more consistent with the average human judgment than human ratings were with one another. However, the lower inter-human agreement, particularly for false memories, suggested that the validation task could be improved. This variability in ratings could be due to inconsistent criteria in how individuals judge true versus false memory. For example, raters might deploy different boundaries between reasonable inferences and inaccuracies based on their prior notions of false memory. We therefore designed a second validation task that was framed as a story comprehension task. Inaccurate recalls were reformulated into two-alternative forced-choice questions. Instead of evaluating subtle distinctions, participants selected the option that best matched their understanding of the story. For example, a factual error "two uncles and two aunts" became: "How many uncles and aunts were there?" with options "two and two" (recalled version) versus "zero and two" (original story). Questions were mostly phrased using wh–constructions (e.g., what, where, when, etc.), and the options were drawn directly from the recalled and original versions of the detail.

For this experiment, we recruited 70 participants, each of whom listened to two stories and completed 25 comprehension trials per story. Compared to the open-ended task, this task substantially improved inter-human agreement (from 0.60 to 0.78 for false memories). Importantly, there was no significant difference between inter-subject and subject–AI agreement in both factual error (subject-AI: 0.78 [95% CI: 0.72, 0.84] vs. inter-subject: 0.81 [95% CI: 0.77, 0.85], p = 0.257) and confabulation trials (subject-AI: 0.67 [95% CI: 0.54, 0.80] vs. inter-subject: 0.75 [95% CI: 0.68, 0.82], p = 0.188, Figure 2B), suggesting that the AI's ratings remained consistent with average human comprehension, even when evaluated under more constrained and objective task demands.

Characteristics of spontaneously occurring false memory

Our validation experiments show that the LLM-based annotation method is an effective way to detect and label memory distortions in our paradigm. Having validated the annotation approach, we next characterized the prevalence and distribution of each false memory type across the entire dataset. We considered a story event as recalled by a participant if at least one recall sentences was matched to that event. An event was labeled as containing a false memory if any matched sentences contained a factual error, confabulation, or both. Importantly, an event could be marked as containing a false detail even if some of its details were accurately remembered–this departs from traditional false memory paradigms in list-learning tasks, in which items are treated as either wholly true or false.

First, we computed the percentage of recalled events which contained a memory distortion for each participant (Figure 2C). Participants produced significantly more factual errors than confabulations across all narratives (factual errors: 0.24±0.01 (SEM); confabulation: 0.09±0.01 (SEM); t = 19.8, $p < 0.001$). For each event, we next calculated the proportion of participants who produced false memories among all participants who recalled that event. We refer to this as the event-level false memory rate. The pairwise correlations for event-level factual error, confabulation and inference rates are shown in Figure S2. Notably, event-level false memory rates varied substantially across the four narratives (Figure 2D), suggesting that story-specific features influence the likelihood of memory distortions. Across all stories, we find that factual errors occurred more frequently than confabulations (factual errors: 0.23±0.01 (SEM); confabulation: 0.08±0.004 (SEM); false memory: 0.28±0.10 (SEM); t = 13.7, $p < 0.001$; Figure 2E).

Semantic factors driving spontaneous false memories

Next we analyzed how semantic features of story events influence the likelihood of false memory. An extensive literature has demonstrated that semantic relationships between studied items are among the strongest predictors of false memory rates[2,42]. In classic paradigms, participants who study semantically coherent word list or view thematically related images exhibit elevated false recognition for category-consistent but non-presented items. This pattern reflects the operation of gist-based memory processes, in which semantic knowledge structures in long-term memory generate plausible but inaccurate details during retrieval. To investigate how semantic context shapes memory distortions in naturalistic recall, we leveraged recent NLP advances[4,21,30,43,44] — specifically, text embedding models 3A-B and autoregressive language models 3C —to quantify semantic features of narratives and test their relationship to false memory types. In our analyses, we used a recent text embedding model (MPNet[45]) that scores highly on a benchmark dataset of human similarity judgments (STS-B[46]) and the GPT-2 language model[47,48]. The correlations for the three semantic predictors are shown in Figure S2.

Semantic centrality

Semantic centrality quantifies the similarity of a given event to all other events in a narrative (Figure 3A). This measure provides a proxy for the importance of any given event in a narrative in conveying the meaning of the narrative as a whole. While recent work using narratives has shown that semantic centrality is a reliable predictor of memory accuracy[4,20], it is unclear how it may influence false memory. Similar to prior work, we constructed undirected graphs representing relationships between narrative events based on their cosine similarity in embedding space.

Similarity to narrative corpus

Prior knowledge structures drive many false memory effects in laboratory paradigms, suggesting that events resembling common world knowledge are more susceptible to memory distortions[42,50–52]. We therefore computed how closely each narrative event matched prototypical narrative content by measuring semantic similarity between individual events and representations derived from a large-scale corpus of diverse stories, NarrativeXL[49] (Figure 3B). This measure is intended to capture the degree to which any given event aligns with stereotypical narrative patterns – for example, common story arcs, character archetypes, or plot conventions that appear frequently across many narratives. Events with high corpus similarity may trigger schema-based false memories because they strongly activate familiar narrative templates, leading to intrusions of expected but non-presented details during recall.

Contextual surprisal

Prediction errors, or surprise, reflect the difference between one's prediction and a real-world output. Surprisal has long been shown to have an influence on memory performance, including for narrative memory[53,54]. Of particular relevance, Sinclair et al.[55] show that unexpected interruptions in videos led to a greater rate of naturally occurring false memories by encouraging the integration of prior knowledge with recently learned information. Therefore, we sought to understand how surprisal may drive naturally occurring false memories in our natural recollections (Figure 3C).

Relating semantic factors to false memory types

We built three separate linear mixed-effects models to predict recall, factual error, or confabulation rate, respectively (Figure 4; Table 1). Participant and story identity were modeled as random intercepts. For fixed effects, we included each of the semantic factors, as well as serial position of the event given its wide-spread influence in list-learning paradigms and possible extension to narrative stimuli[4,29,56]. Finally, we also included event length as a fixed effect, as a memory error is more likely to occur in longer events given the current approach.

This analysis revealed three key findings (Table 1). First, consistent with prior studies[4,20], events with higher semantic centrality were more likely to be recalled ($\beta = 0.10$, $p < 0.001$). There was no significant effect of recall rate with respect to contextual surprisal ($\beta = -0.03$, $p = 0.145$) or similarity to narrative corpus ($\beta = -0.03$, $p = 0.189$). Second, we found a strikingly clear double-dissociation between semantic factors and the two types of false memory. Events that closely resembled the narrative corpus contributed to more factual errors but not confabulations($\beta = 0.14$, $p = 0.001$), while surprising events

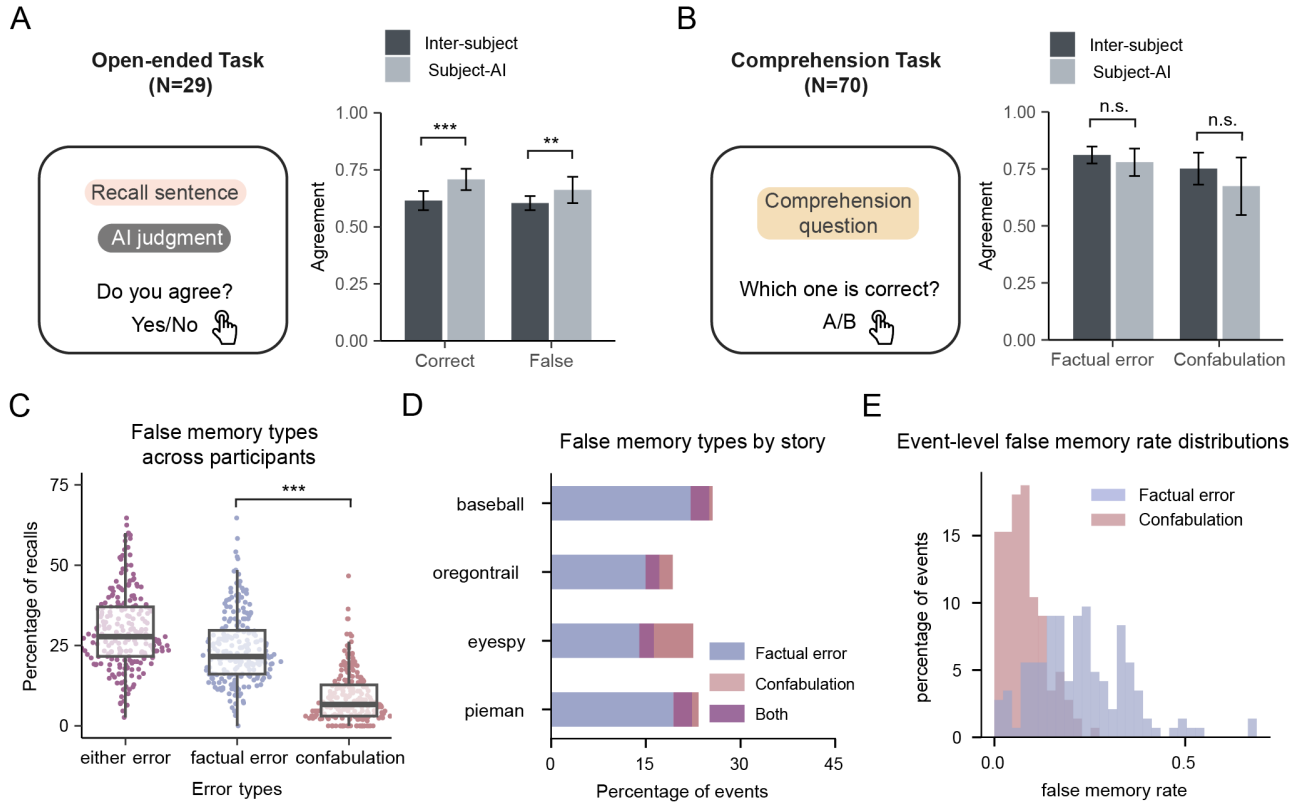

Figure 2. Results from human validation experiments and summary of GPT-4o classification of false memories. **(A)** Subject-AI agreement exceeded inter-subject agreement in Validation Experiment 1. Twenty-nine raters evaluated 60 recall sentences (30 per story), each shown with story context and the AI's binary classification (accurate vs. inaccurate), with incorrect segments underlined in false memory cases. Raters made a binary AI agreement judgment and rated their confidence. Subject–AI agreement (light gray) was significantly higher than inter-subject agreement (dark gray) for both accurate, or correct trials ($p < 0.001$) and false memory trials ($p = 0.002$)–suggesting the AI's judgments better reflected the group consensus. (*** $p < 0.001$; ** $p < 0.01$). **(B)** AI ratings aligned with group-level comprehension in Validation Experiment 2. To reduce ambiguity, we developed a structured comprehension task where seventy raters completed 50 questions (25 per story). Inaccurate recall segments were converted into two-alternative forced-choice questions (falsely recalled detail vs. true original detail). This format improved inter-subject agreement (0.60 to 0.78 across memory types). No significant differences emerged between subject–AI and inter-subject agreement for either factual errors ($p = 0.257$) or confabulations ($p = 0.188$). For A-B, error bars represent bootstrapped confidence intervals. **(C)** Average percentage of recalled events containing memory distortions per participant (each dot in swarm plot corresponds to a single participant). Participants exhibited a higher proportion of factual errors (0.24 ± 0.01 SEM) compared to confabulations (0.09 ± 0.01 SEM; $t = 19.8$, $p < 0.001$). **(D)** Proportion of participants who falsely recalled each event, averaged across events within each story. Bar colors indicate the proportion attributed to factual error (blue), confabulation (red), or both (purple). Variability across narratives suggest that their content influences the rate of false recall. **(E)** Histogram of the proportion of events with different false memory rates, based on values from (D). Overlaid histograms separately show the distribution for factual errors (blue) and confabulations (red). For a given event, confabulations were less frequent than factual errors (left skewed distribution), consistent with the view of the data from (C).

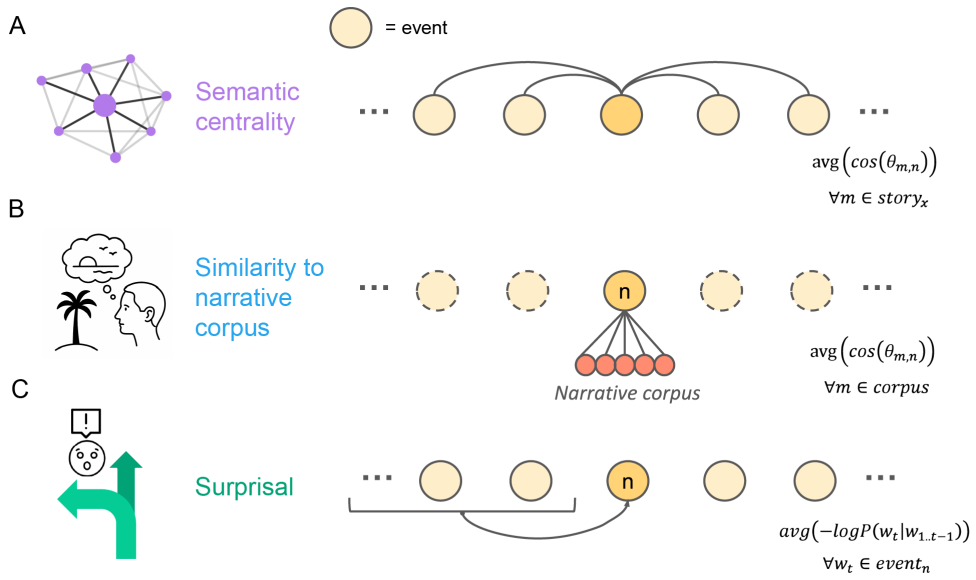

Figure 3. Computational measures of semantic context in narrative events. **(A)** Semantic centrality quantifies how well-connected each event is to the overall narrative structure. Event text is embedded in high-dimensional semantic space, and centrality is calculated as the average cosine similarity between each event and all other events within the same story[4,20]. **(B)** Similarity to narrative corpus measures how closely each event matches semantic patterns in a selected text corpus, such as a corpus of narratives. This is measured as the average similarity of a given event embedding to all events in the corpus[49]. **(C)** Contextual surprisal captures how unexpected each event is given the preceding narrative context. Surprisal is computed as the negative log probability of each word conditioned on prior context, averaged across all words within the event.

were associated with more confabulations ($\beta = 0.16$, $p = 0.002$) but not factual errors. Semantic centrality did not contribute significantly to either false memory metrics. Third, we found serial position to negatively contribute to factual error rate ($\beta = -0.16$, $p < 0.001$), but not recall rate ($\beta = -0.01$, $p = 0.477$) or confabulations ($\beta = 0.08$, $p = 056$). That is, events later in the story had fewer memory distortions. The full results from the mixed effects models are shown shown in Table 1.

To test the robustness of these results, we re-computed semantic factors using an alternative embedding model (Universal Sentence Encoder[57]) and an alternative language model for computing surprise (Pythia[58]). The choice of these models is based on existing benchmarks or prior behavioral modeling studies (see LLM-based Annotation Pipeline in Methods section). We replicated the disassociation between factual error (USE and Pythia: centrality: $\beta = -0.13$, $p < 0.001$; similarity to narrative corpus: $\beta = 0.15$, $p < 0.001$) and confabulation factors (contextual surprisal: $\beta = 0.12$, $p = 0.028$), as well as the effect of centrality on recall rate ($\beta = 0.9$, $p < 0.001$). See Table S5 for linear mixed effects models' results using USE and GPT2-Small, and Table S6 for those using USE and Pythia.

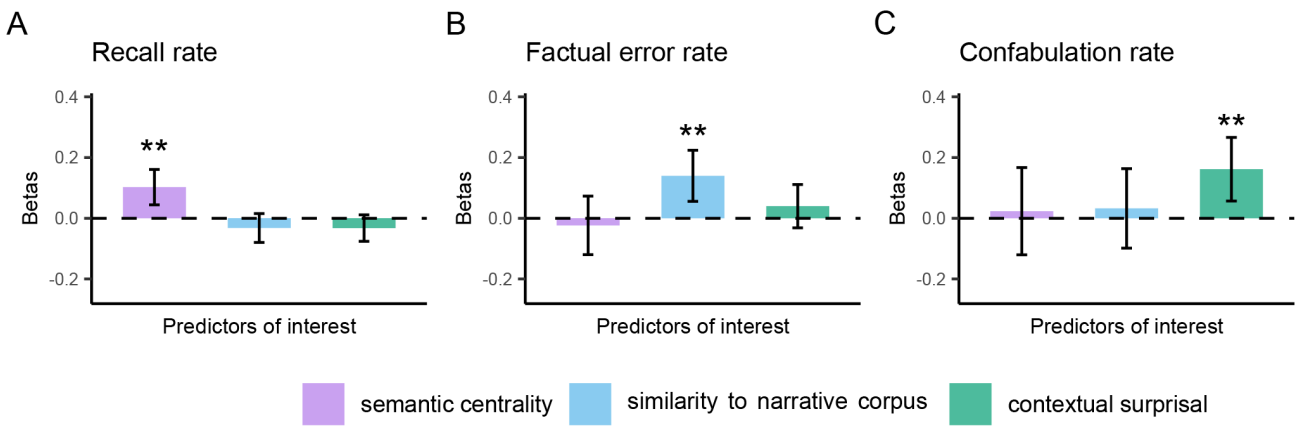

Figure 4. Semantic factors differentially predict memory accuracy and distortion types. Linear mixed-effects models were used to assess how semantic properties of story events predict recall likelihood and different forms of memory distortions. **(A)** Events with higher semantic centrality are more likely to be recalled, consistent with the importance of this measure for memory retention[20] ($\beta = 0.10$, $p < 0.001$). **(B)** Factual errors are more likely to occur when events have greater similarity to narrative corpus ($\beta = 0.14$, $p = 0.001$). **(C)** Confabulations were more frequent for events with higher contextual surprisal ($\beta = 0.16$, $p = 0.002$), suggesting a link between unpredictability and memory fabrication. Bars show standardized regression coefficients (β); error bars indicate 95% confidence intervals. All models included semantic centrality, similarity to narrative corpus, surprisal, serial position, and length as fixed effects, with participant and story identity modeled as random effects (see Table 1).

	β **coefficient**	**CI (Wald)**	p **value**
Recall: 229 participants (16,345 trials)			
Intercept	0.09	–0.33, 0.51	0.674
Centrality	0.10	0.04, 0.16	**<0.001**
Contextual surprisal	–0.03	–0.08, 0.01	0.145
Sim. to prior corpus	–0.03	–0.08, 0.02	0.189
Event length	0.32	0.28, 0.36	**<0.001**
Serial position	–0.01	–0.05, 0.02	0.477
Factual error: 229 participants (8,295 trials)			
Intercept	–1.21	–1.45, –0.97	**<0.001**
Centrality	–0.02	–0.12, 0.07	0.635
Contextual surprisal	0.04	–0.03, 0.11	0.273
Sim. to prior corpus	0.14	0.06, 0.22	**0.001**
Event length	0.10	0.05, 0.16	**<0.001**
Serial position	–0.16	–0.21, –0.10	**<0.001**
Confabulation: 229 participants (8,295 trials)			
Intercept	–2.63	–2.85, –2.41	**<0.001**
Centrality	0.02	–0.12, 0.17	0.752
Contextual surprisal	0.16	0.06, 0.27	**0.002**
Sim. to prior corpus	0.03	–0.10, 0.16	0.627
Event length	0.13	0.05, 0.22	**0.002**
Serial position	0.08	0.00, 0.17	0.056

Table 1. Linear mixed-effects model results for predicting recall, factual error, and confabulation rates. Standardized regression coefficients (β), 95% Wald confidence intervals (CIs), and p-values are reported for each fixed effect in the three models. The significance threshold was Bonferroni-corrected for multiple comparisons ($\alpha = 0.0167$). p-values for significant predictors are shown in **bold**. All models included participant and story identity as random intercepts.

Discussion

This work establishes a computational approach for detecting memory distortions in naturalistic narrative recall, revealing that distinct semantic factors drive separable classes of false memories. We demonstrate that contextual surprise predicts confabulations—entirely fabricated details—while similarity to prototypical narrative patterns predicts factual errors—distortions of narrative content. This dissociation provides the first empirical evidence that factual errors and confabulations reflect distinct cognitive mechanisms operating during memory reconstruction, supporting theoretical proposals that have lacked direct empirical validation in naturalistic contexts[37,38,59,60]. Critically, this separation emerges in naturalistic recall, where participants generate rich, unconstrained responses - a distinction that would be impossible to detect using traditional list-learning or recognition paradigms that treat false memories as a unitary phenomenon. Our LLM-based approach enables scalable analysis of these complex memory processes without requiring extensive manual annotation, opening new avenues for studying how semantic context shapes memory errors in real-world settings.

Lessons learned from validating model performance

Validating LLM performance in classifying psychological phenomena presents significant methodological challenges. We validated model classifications of memory distortions through direct comparison with human ratings on a subset of participants' recollections. In our open-ended experiment, we found that human inter-rater reliability was notably low (60% for false memory trials), despite participants receiving detailed instructions and access to the complete story text and relevant narrative context. Importantly, human-AI agreement was significantly higher (66% for false memory trials), suggesting that the AI's classifications more closely approximated the average human judgment than individual human raters agreed with each other, which was observed in other psychological tasks comparing AI task behaviors and human raters[27,61].

This low inter-rater agreement reveals a fundamental challenge in validating computational approaches to psychology: many constructs lack objective ground truth[39]. Individual differences in interpreting what constitutes "false" memory in naturalistic recollections may reflect genuine variability in how humans process and evaluate memory distortions. This challenge extends beyond false memory research to the validation of many psychological constructs in naturalistic settings. Studies of event segmentation[27,62–64], causal structure[65,66], and narrative coherence[4,21] have similarly relied on aggregated human judgment to establish statistical ground truth when objective criteria are unavailable or inadequately defined.

To test whether task ambiguity contributed to low inter-rater reliability, we conducted a second validation experiment that reframed memory evaluation as a multiple-choice comprehension task with clearly defined alternatives. This design eliminated the need for participants to make subjective judgments about memory accuracy by instead asking them to select which version of a detail—the original story content or the participant's recalled version—better matched their understanding of the narrative. This approach substantially improved inter-rater reliability (78% for factual errors, 75% for confabulations) while maintaining comparable human-AI agreement, confirming that our computational approach remained aligned with human comprehension even under more constrained evaluation conditions.

These validation experiments demonstrate that our LLM-based approach provides reliable detection and classification of memory distortions in naturalistic recall. More broadly, our findings suggest that future computational studies of psychological phenomena may benefit from validation designs that minimize interpretive ambiguity, either through objective task framing or by explicitly acknowledging and modeling human variability as an inherent feature of the construct under investigation.

Semantic metrics drive distinct memory distortions

A central contribution of this work is demonstrating that different semantic factors drive distinct types of memory distortion. Our findings reveal a clear dissociation: events resembling prototypical narrative content generate factual errors—distortions of actual story details—while semantically surprising events trigger confabulations—entirely fabricated information. This dissociation provides direct empirical evidence for theoretical proposals that false memories reflect multiple underlying mechanisms operating during memory reconstruction[37,38,59,67].

The role of prior knowledge in generating factual errors aligns with established findings from controlled paradigms. In DRM tasks, semantic associations between studied and non-studied items produce robust false recognition[2,42,50,68], while misinformation studies demonstrate how existing beliefs shape memory distortions[3,11,12]. Our results extend these findings to naturalistic settings, showing that narrative events matching familiar schemas are particularly vulnerable to factual errors. This suggests that when story content strongly activates existing knowledge structures, memory reconstruction may inappropriately integrate expected but non-presented details.

Our findings suggest that the link between contextual surprise and confabulations may reflect a distinct cognitive process. Surprising events violate predictions based on preceding narrative context, and may trigger compensatory mechanisms that generate novel content to maintain narrative coherence. This finding converges with research on flashbulb memories, where unexpected real-world events show heightened distortion rates[69,70], as well as recent work demonstrating that unexpected narrative breaks increase false memory occurrence[55]. The fabrication of entirely new details following surprising events may

represent an adaptive response to prediction error, in which the memory system attempts to resolve contextual inconsistencies during encoding.

Our findings provide new evidence of a functional separation between processes that drive false memory. While theoretical frameworks have proposed frameworks with distinct types of false memories, empirical studies typically use single paradigms to study single types[32,37,60,67,71–73]. Others have noted that the empirical validation of distinct false memory types has remained elusive due to methodological constraints[74]. Traditional paradigms typically classify recall as simply correct or incorrect, precluding fine-grained analysis of error types[17]. Our naturalistic approach reveals qualitative differences in memory distortion that emerge only when participants generate rich, unconstrained responses reflecting the full complexity of reconstructive memory processes.

The ecological validity of narrative recall proves essential for detecting this dissociation. Unlike list-learning or recognition tasks that constrain possible responses, free recall of complex narratives allows multiple types of memory errors to co-occur and compete, creating the behavioral richness necessary to distinguish between factual errors and confabulations. This methodological advance opens new avenues for investigating how different cognitive mechanisms contribute to memory distortion in real-world contexts, where multiple semantic factors interact dynamically during retrieval.

Our findings offer a key contribution by linking specific semantic features to distinct types of false memory. We replicate prior work showing that surprisal and prior knowledge influence memory errors, but extend it by demonstrating a clear dissociation: factual errors are predicted by similarity to narrative corpus, while confabulations arise from surprisal. This pattern aligns with prior studies showing that memory of surprising real-world events, particularly flashbulb memories, has a high likelihood of distortion[69,70], and that unexpected breaks in narrative continuity increase false memory rates[55].

Limitations

In the current study, we demonstrate that narrative memory distortions can be reliably identified and meaningfully analyzed using computational approaches at scale. Across studies and participants, false memories occurred in up to 27.6% of recalled events, with factual errors (22.6%) more frequent than confabulations (7.7%). Building on established uses of generative language models in memory research, we made key innovations with the use of in-context learning examples and a novel taxonomy for classifying memory errors.

The study has several limitations which should be addressed. First, we determined the two types of false memories apriori based on existing literature, but may have overlooked other error types, especially given the complexity of naturalistic recall. Future work should explore the interactions of more known memory error types, and how individuals differ in drawing boundaries between them. Second, while we provide evidence for a double-dissociation and hypothesize about why these processes are distinct, further research will be needed to test these hypotheses in naturalistic recall.

Real-world implications

False memory research carries important real-world consequences, particularly in domains such as eyewitness testimony where errors can shape legal outcomes[75,76]. In this work, we refrain from assigning harm or deception to the term "false memory"; rather, we use it broadly to describe distortions that can arise in recall that can encompass a range of types[17,77,78]. The prevalence rates we report should therefore be interpreted with caution: they reflect specific experimental choices and should not be taken as general benchmarks of false memory occurrence. Similarly, our automated detection method is not intended for judging the truth of individual statements. Its value lies in revealing variability across individuals and events to study the systematic conditions leading to memory distortions. The classification methods used in this study serve as research tools for understanding memory phenomena, not as instruments for evaluating truth or accuracy in real-world situations.

Methods

Free recall experiment

Participants

We used the full cohort from our recently published dataset (the "Naturalistic Free Recall" dataset; NFRD[4]). This includes 229 native English speaking participants (145 female; $Mean_{age}$ = 25.03, SD_{age} = 11.15), who took part in the online study (see Table S1 for detailed demographic information). This cohort was selected after excluding participants based on self-reported engagement (lower than or equal to 2 out of a 5-point Likert scale). Data collection was split between New York University's SONA Systems platform (n = 167; $Mean_{age}$ = 19.67, SD_{age} = 1.33), where undergraduates earned course credit, and Prolific (www.prolific.com; n = 62; $mean_{age}$ = 39.77, SD_{age} = 12.90), whose members were compensated at a rate of $10 per hour. All participants provided informed consent prior to experimental testing. This study was approved by New York University's Committee on Activities Involving Human Subjects (IRB-FY2016-1357).

Materials

The same four spoken narratives used in the NFRD were leveraged in this study. The set comprised three personal stories from The Moth Radio Hour—"Pieman", "Eyespy" and "Oregontrail"—alongside the first chapter of Lester Chadwick's "Baseball Joe in the Big League" (hereafter "Baseball"), obtained from LibriVox (www.librivox.org). Each audio file was paired with a verbatim transcript: the "Pieman" text was drawn from a previously published neuroimaging dataset[53], and the "Baseball" transcript was sourced from Project Gutenberg (www.gutenberg.org); for "Eyespy" and "Oregontrail,"[4] generated and verified the narrative transcripts. The NFRD also segmented the narrative transcript text into events, using a previously validated method[21], with some optimization in parameter selection[4] . On average, each audio clip lasted 11 minutes and 35 seconds. A complete overview of the stimuli and their sources is provided in Table S1.

Participant verbal recollections were first transcribed using the Speech-to-Text API offered by Google Cloud (cloud.google.com/speech-to-text), then proofread and corrected by a professional transcription service to ensure high-fidelity transcriptions (Figure 1A;[4] for complete details).

LLM-based Annotation Pipeline

We developed a two-stage automated classification pipeline (Figure 1B) using GPT-4o (version "2024-05-13")[31] to identify and categorize false memories in text recollections. Each stage was implemented as a prompt-driven interaction using LangChain (www.langchain.com) with Python 3.9. The entire text of the original story was provided in the prompt (see below, Tables S2, S3 and S4). To perform the scoring, each recall sentence was given in a new conversation message. To stay below context window limits, we maintained a conversation history with the five most recent recall sentences and AI annotations. Model temperature was set to 0.3 for event alignment and 0 for memory classification to balance between task reproducibility and flexibility, a practice based on prior literature[79,80].

Event matching task In the first stage, the LLM was prompted to align each recall sentence with a corresponding narrative event (Figure 1B). The system prompt consisted of two key components (Table S4): (1) instructions directing the model to match each recall sentence to one of the numbered story events, and (2) the full list of story events, each labeled with an ordered index.

Recall transcripts were segmented into individual sentences. Sentences shorter than 8 characters were concatenated to the preceding sentence to ensure sufficient content for event matching. For each transcript, a new GPT-4o[31] instance was initialized with the system prompt, and recall sentences were passed in sequentially. The model's index outputs (e.g., "event 5" or "5") were recorded for subsequent analysis. Outputs that are either index "0" or texts without (e.g., "I'm sorry, but I need more context to identify the specific event. ") indices were coded as no match for the recall sentence.

In contrast to prior work[4], our method uses sentence-level granularity to better align with generative LLMs, which process grammatically complete content more effectively. Event-level recall, defined as the proportion of participants with at least one sentence matched to an event, strongly correlated with prior recall measures using event-level granularity in participants' recalls ([4]; $r = 0.46$, $p < 0.001$).

Memory classification task

In the second stage, the LLM was prompted to classify each recall segment into different types of memory distortions. This step used in-context learning, leveraging the model's ability to perform complex tasks based on task-specific examples without updating model parameters[81]. In-context learning is particularly suitable here because it allows the model to flexibly adapt to nuanced psychological categories[82]. To ensure the model interprets the task as psychological research and responds appropriately, the role of the human speaker in the conversation chain was renamed from "user" to "psychologist".

The system prompt included three components (see Table S2 and S3 for full content): (1) detailed classification instructions, (2) the complete original story for reference, and (3) eight in-context examples illustrating classification labels.

Our classification framework contained three mutually exclusive categories for memory distortions (Figure 1C): factual errors (details that contradict the original story), confabulations (fabricated details not present in the story), and inferences (plausible but unmentioned details derived from prior knowledge or reasonable interpretations). These types of memory distortions, including inferences[34], were inspired by theoretical accounts[38,60] and empirical studies on memory distortions[74].

The in-context examples included two representative cases of each memory type, i.e., factual error, confabulation, and inference, as well as two examples that did not fall into any of these categories. This balanced set was designed to guide model classification decisions and reduce ambiguity for edge cases (see Table S3). After promoting the model, recall segments were given to the model in sequential order. For each recall segment, the model returned labeled text outputs for each of the three categories (see Table S3 for in-context learning examples). The LLM was instructed to either return the verbatim portion corresponding to the memory distortion, or to return "None" to indicate that the memory distortion did not appear in that recall segment (Figure 1C).

Validation Experiment 1 (open-ended task)

Participants

We recruited 30 participants through the NYU's SONA subject pool to evaluate the alignment between GPT-4o's false memory classifications and human judgments. To disambiguate between participants in the validation task and the original recall task, we henceforth refer to validation task participants as raters. Raters were randomly assigned to one of two story pair conditions: 13 were assigned to Pieman and Eyespy, and 17 to Oregontrail and Baseball. One rater in the Pieman/Eyespy group was excluded for selecting "neutral" on the item "I understood the task instructions," leaving a final sample of 29 raters (12 for Pieman/Eyespy and 17 for Oregontrail/Baseball; $mean_{age}$ = 18.61, SD_{age} = 1.06; 17 female). All raters were native-english speakers and completed the task in person.

Materials

To validate the AI-generated false memory labels, we selected a balanced and representative sample of 30 recall sentences per story (120 sentences total). This was done by randomly selecting six events from each narrative. As such, this approach was intended to avoid potential selection bias and allowed us to test the model's performance across a range of stories and events. For each event, we randomly sampled five recall sentences that had been matched to that event by our annotation pipeline. This ensured that multiple participants' recollections for the same narrative were evaluated. To balance the evaluation and reduce bias in rater judgments, we selected an equal number of recall sentences labeled by the AI as accurate and as containing a memory distortion (15 per each story; 78.1% are factual errors and 21.9% are confabulations across all stories). The resulting 30 trials per story provided a manageable yet statistically meaningful sample. Each rater reviewed two stories (60 trials total).

Task

We programmed the task using JsPsych[83] and hosted it on Google Firebase (https://firebase.google.com/). At the beginning of the in-person experiment, the experimenter introduced the purpose of the study, namely to provide a separate evaluation of false memory for recalled details of a story, given an existing rater's judgment. Notably, human raters were not told that they were evaluating an AI rater. The experimenter then walked each rater through sample trials to ensure they understood the instructions and task format. Raters then proceeded to the main task on lab computers. The task was divided into two blocks, one for each of their assigned stories.

In each block, raters listened to the full audio recording of the assigned story. This ensured that they were familiar with the story content and could make informed judgments during the validation trials. The raters were also provided with a physical print-out of the story transcripts for reference during the task. Following the listening phase, they were presented with a series of validation questions, each corresponding to a single recall sentence that had been previously matched to a specific story event.

Each trial displayed the matched story context on the left side of the screen and the corresponding participant recall on the right side. The story text on the left consisted of the focal event along with the immediately preceding and following events for context. This consistent presentation was intended to help raters locate the relevant section of the narrative while encouraging them to draw upon their broader understanding of the story. The trials were presented in mini-blocks: a single story event was given and raters viewed and validated all five recalls for that event during the mini-block. Raters were instructed to try to base their judgments on the entire story, not just the story context given on the left side of the screen.

At the top of the screen, raters viewed the AI model's binary classification for the recall sentence: either "The rater thinks this recollection is accurate" for correctly classified sentences, or "The rater thinks this recollection contains inaccuracies" for those containing false memories. For trials involving false memories, the recall text that was returned by the model was underlined to guide the rater evaluations.

The task required the raters to agree or disagree with the AI's classification (two-alternative forced choice), followed by a 5-point Likert confidence rating ranging from "not at all" to "completely" (labels: "not at all", "slightly", "somewhat", "fairly", "completely").

After completing the experimental blocks, raters answered a short post-task questionnaire that assessed their engagement and experience. This questionnaire included statements such as "I understood the task instructions," "I was engaged in the experiment," and "I found the experiment difficult," all rated on a 5-point Likert scale. These responses were used to assess data quality and to confirm raters' understanding and task compliance.

Analysis

To evaluate how reliably humans aligned with the automated pipeline, we computed two agreement measures: inter-subject agreement and subject–AI agreement. Agreement refers to the extent to which two raters (human-human or human-AI), made the same binary judgment about a recall trial (i.e., whether they both judged the recall as accurate or inaccurate). We computed agreement as a simple proportion (e.g., matched judgments on 55/60 trials). By comparing inter-subject agreement and subject-AI agreement in a permutation test, we test whether AI provides a reasonable proxy for human judgment[25,27,40,41].

For inter-subject agreement, within each story, we calculated the proportion of identical responses for every possible pair of raters across all trials[28]. These pairwise agreement scores were averaged to yield a single inter-subject agreement score per story. For subject–AI agreement, we computed the proportion of trials on which each rater's response matched the AI's original classification and averaged those proportions across raters within each story.

To estimate uncertainty and compare the agreement metrics statistically, we required a structured format that captured rater-level judgments across trials. To this end, we constructed two binary matrices per story: one representing subject–AI agreement and another representing inter-subject agreement. The first matrix (the dimension is the number of questions by the number of raters) indicated whether each rater agreed with the AI on each trial (1 for agreement, 0 for disagreement). The second matrix (the dimensions is the number of questions by the number of all possible rater pairs) indicated whether each pair of raters agreed with each other on each trial. These matrices provided a structured format for quantifying trial-level agreement patterns and laid the groundwork for further statistical analysis.

We used nonparametric bootstrapping to estimate confidence intervals and test for statistical differences between agreement types. For confidence intervals, we resampled the trial rows of each matrix with replacement 2,000 times per story, recalculating average agreement for each resample to generate 95% confidence intervals. To aggregate results across stories, we computed pooled standard errors by taking the square root of the sum of squared standard errors divided by the number of narratives.

To test whether subject–AI agreement exceeded inter-subject agreement, we employed a two-stage bootstrap procedure to account for both within-story and between-story variability. First, we resampled trial rows 2,000 times per story and calculated the difference between subject–AI and inter-subject agreement for each resample, yielding a bootstrap distribution of agreement differences for each story. Second, we performed 2,000 iterations where we randomly sampled (with replacement) one difference value from each story's distribution and averaged them, generating a grand bootstrap distribution of agreement differences across stories. From the bootstrap distribution of averaged effects we obtained 95% confidence intervals as the 2.5th and 97.5th percentiles, and calculated two-sided p-values by doubling the fraction of bootstrap means that lie on the opposite side of zero from the observed mean.

Model validation Experiment 2 (comprehension task)

Participants

For the second Validation Experiment, we recruited 126 raters using NYU's SONA system. Raters were given course credit for their participation. We applied two exclusion criteria to ensure data quality. First, same as Validation Experiment 1, we removed 39 raters based on their self-reported engagement and level of understanding of the task (lower than or equal to 3 out of a 5-point Likert scale). Second, we excluded 17 raters who scored below 90% on embedded engagement check questions. These checks were designed to identify individuals who had not adequately understood the core story content. The final sample included 70 raters ($mean_{age}$ = 20.13, SD_{age} = 1.14; 41 female). Raters were randomly assigned to receive two of the four stories. The number of unique raters completing each story was as follows: Pieman (47), Eyespy (39), Oregon Trail (29), and Baseball (25).

Materials

This experiment adapted the materials from Validation Experiment 1 by adopting a more constrained comprehension-based format. In addition to the trials used in Validation Experiment, we selected four other events (i.e., 20 recall sentences), yielding 50 recall sentences per story. We focused exclusively on half of those that the AI had classified as containing false memories (either factual errors or confabulations), as these were the primary targets for validation.

The sentences were reformulated into a two-alternative forced-choice comprehension question using either wh- (who, what, where, etc.) or yes–no formats. Questions focused on the specific detail that the AI had identified as inaccurate. We aimed to preserve the original sentence structure while isolating the distorted element. The AI-identified erroneous phrase became one answer option (the "recalled" option), while the corresponding correct detail from the original story served as the alternative (the "verbatim" option). For example, a recall sentence containing the incorrect detail "he asked the principal..." was converted to: "Who did Jim question about whether the raise is against the school's policy?" with options "the dean" (verbatim) and "the principal" (recalled). We also included yes/no questions when the verbatim detail was unclear or unmentioned. For example, the recall sentence "He went to the dean's office" was converted to "Did Jim go to Dean McGowan's office?".

To assess raters' overall task engagement, we also included five engagement check questions per story (catch trials). These questions were used to identify raters who failed to grasp the core narrative content. We randomly sampled five events from the existing question set and added one extra question asking about basic factual details of the event. The two choices were a verbatim answer and an incorrect answer that we designed to be very easily identified. For example, one engagement question in the story Pieman was "Did the person who pied the dean run away?" with the options yes and no.

Each rater was randomly assigned two out of the four stories, receiving all 25 comprehension questions and five engagement checks per story.

Task

Raters completed the experiment remotely using their personal devices. Following the design of Validation Experiment 1, raters listened to two complete story recordings and then completed the associated comprehension trials for each story.

Each trial presented relevant story context consisting of the focal event and its immediately preceding and following events. Raters were instructed to try to base their responses on their complete understanding of the story, not just on the local context provided. Furthermore, raters were instructed to select one of the two options before advancing to each subsequent trial. Upon completing all trials, raters completed the same engagement questionnaire from Experiment 1 to assess task understanding and effort.

Analysis

We evaluated agreement for each question containing a memory error. Raters were considered to agree with the AI if they selected the verbatim option, which aligned with the detail from the original story. Selecting the recalled (i.e., false memory) option was treated as disagreement with the AI's judgment.

We then computed inter-subject and subject–AI agreement separately for questions derived from factual errors and confabulations, respectively. To test for significant differences, we applied the same two-stage nonparametric bootstrapping procedure used in Validation Experiment 1.

Evaluating the distribution of false memories

To investigate how false memories manifest across participants and story events, we combined outputs from both the event matching and memory classification tasks. For each story event at the participant-level, we determined: (1) whether the event was recalled—defined as at least one recall sentence being matched to that event by the event alignment task—and (2) whether any of those matched recall sentences contained a factual error, confabulation, or inference, as identified by the classification task.

If a participant recalled a story event and any of the corresponding recall sentences included a false memory (i.e., either a factual error or a confabulation), we considered that participant to have recalled the event with a distortion. This approach ensured that multiple recall attempts for the same event were consolidated at the event level, and that the presence of any memory distortion was sufficient to count the event as distorted for that participant. Prior to all subsequent analyses, we excluded story events that were recalled by fewer than 20 participants. This threshold was chosen to ensure sufficient sample size for estimating event-level false memory rates.

Using this method, we computed two sets of measures. At the participant level, we calculated the proportion of recalled events that contained each type of distortion. At the event level, we calculated the proportion of participants who recalled each event with an error, normalized by the total number of participants who recalled that event at all. This allowed us to quantify the rate of false memories for any given event across participants.

To compare the relative rates of each error type, we conducted paired t-tests on both participant-level and event-level data, comparing the rates of factual errors and confabulations.

Extraction of semantic features

To characterize the semantic properties for individual events, we extracted three event-level features: semantic centrality, similarity to narrative corpus, and contextual surprisal.

Our primary analyses relied on sentence embeddings and perplexity (as a measure of model-based surprisal) values derived from pretrained embedding models and autoregressive language models, respectively. For embeddings, we used the `all-mpnet-base-v2` model[45] via the Sentence Transformers Python library[84], chosen for its strong performance on the STS-B benchmark of human similarity judgments[46]. For surprisal estimation, we employed GPT-2 Small (124M parameters)[47,48], the latest open-source GPT model demonstrated to predict human reading and memory behaviors[53,85–88].

To assess the robustness of our findings, we also tested an alternative embedding model —`Universal Sentence Encoder` (USE)[57] and one alternative autoregressive language model of similar size: Pythia (160M)[58]. USE was selected for its broad applicability in narrative memory studies[4,20,89–91]. We chose Pythia because it has a similar number of parameters to GPT-2 small, and has also been shown to predict human reading times[92]. Results of the linear mixed effects models from these alternatives are reported in the supplementary materials. All models were downloaded from Hugging Face (`huggingface.co`).

To compute semantic centrality, we calculated the pairwise cosine similarity between the embedding of each event and all other events within the same story (in line with previous work[4,20]). We then averaged these similarity scores for each event, yielding a measure of how semantically connected an event was to the rest of the narrative—an index of its conceptual centrality within the story's overall structure.

To estimate similarity to prior narrative corpus, we compared each story event to a large external corpus of fiction using the NarrativeXL dataset[49], which includes full-text versions of 1,500 fiction books from Project Gutenberg. We segmented into

55-word windows to match the structure of our narrative events, resulting in approximately 2.3 million segments comprising over 126 million words. We embedded each segment using the same sentence embedding model and computed the average cosine similarity between each story event and all NarrativeXL segments. This score quantifies how closely each event resembles prototypical narrative content—that is, common story patterns found across a large fiction corpus.

As a measure of contextual surprisal, we estimated how suprising each event was given its preceding story context. Using GPT-2 Small, we concatenated each event with all preceding words in the story and computed token-level cross-entropy loss. A binary mask was applied to isolate the tokens from the current event, and the average negative log-likelihood (NLL) over those tokens was used as the surprisal score. Higher values indicate greater deviation from expected narrative progression. The analysis was implemented using PyTorch[93], and Hugging Face's Transformers library[94]. Results from an alternative comparable model (Pythia[58]) are provided in the supplementary materials for comparison (Table S5 and S6).

Relating semantic measures to memory behavior

To investigate how semantic features influence both memory accuracy and error rates, we constructed three linear mixed-effects models predicting participants' behavioral responses. The models targeted three distinct outcomes: (1) whether an event was recalled, (2) whether it was recalled with a factual error, (3) whether it was recalled with a confabulation. Note that the first model included unrecalled events, whereas the second and third models only contained recalled events. These models allowed us to test which event-level features predicted successful recall and specific types of memory distortion. All models were fit using the lme4 package (version 1.1-28) in R (version 4.1.2)[95].

We selected a consistent set of fixed effects across models. These included: semantic centrality, similarity to prior narrative corpus, contextual surprisal, serial position (order of the event within the story), and event length (in words). All predictors were z-score normalized prior to model fitting. Participant and story ID were included as random intercepts to account for individual differences and story-level effects.

To control for multiple comparisons across the three memory error models, we applied Bonferroni correction, adjusting the significance threshold by a factor of three ($\alpha = 0.0167$). This conservative approach reduces the likelihood of Type I errors when testing related hypotheses across the recall, factual error, and confabulation models. The significance values were generated as part of the summary output based on asymptotic Wald tests.

Code and data availability

All of our analysis code are available at https://github.com/phoebsc/Distinct-paths-to-false-memory. The published narrative recall dataset[4] analyzed in the study can be downloaded at https://osf.io/h2pkv/.

Supplementary Information

	Stimulus source	Speaker gender	Length in words	Length in audio (seconds)	Participant demographics		
					N	age range	gender
pieman	The Moth Radio Hour	Male	948	489	116	17-29	88 female
eyespy	The Moth Radio Hour	Female	2318	779	116	17-29	88 female
oregontrail	The Moth Radio Hour	Female	2389	743	113	18-75	57 female
baseball	LibriVox Audio	Male	2088	768	113	18-75	57 female

Table S1. Overview of the narrative stimuli and participant demographics.

	System prompt content
Instruction	You are going to hear a story told by [name of the narrator] in its entirety, and then see a subject's recall of the story sentence by sentence. Using common sense, you will rate whether each recall sentence contains two mutually exclusive types of false memories. Please note paraphrasing or transcription typos are not false memory. False memory that is corrected by the subject themself does not count, either. 1) *Factual error*: Where in the statement, if any, contains factual error that contradicts the story? Factual errors include wrong subject, object, location, timing, characters' speech or action. 2) *Made-up information*: Where in the statement, if any, contains baseless information absent in the story? This includes new characters or actions that are intuitively false and not inferences or factual errors. In addition, you will also include ratings for: 3) *Inference*: Where in the statement, if any, has inferences derived from common world knowledge or characters' mental state?
Story text	[Full story transcript]
Examples	Eight examples in the following format: Psychologist: Answer: 1) [details considered factual errors or "None"] 2) [details considered confabulations or "None"] 3) [details considered inferences or "None"]

Table S2. System prompt template for the memory type classification model. The system prompt consists of three sections: instruction, the full story transcript, and in-context learning examples. The exact content varies based on the story.

	In-context learning examples
Pieman	Psychologist: "He recalls the place, but I don't quite remember it. Rhode Island in Massachusetts? I forgot." answer: 1) "Rhode Island in Massachusetts?" 2) None. 3) None. Psychologist: "and so the guy gets really excited, and he goes and starts writing the story up. It's his first-ever story. He's really excited, and he writes a story up but better." answer: 1) None. 2) "It's his first-ever story." 3) None. Psychologist: "which is what Pie Man would say after an attack, or like for justice. An attack for justice, not in a negative connotation." answer: 1) None. 2) None. 3) "An attack for justice" Psychologist: "The guy ends up going to those library steps at that time and sees the student body president, who's female. I forgot her name. She is, as the guy describes, well-bred and is not working class like most of the foreign student body was, and is wealthy" answer: 1) None. 2) None. 3) "is not working class like most of the foreign student body was, and is wealthy"

Continued on next page

	In-context learning examples
	Psychologist: "Then he goes to a bar one night and meets a girl named Amanda I believe, and he has a crush on her or he says like they were flirting, but then he said he was actually the only one flirting with her." answer: 1) "Amanda" 2) None. 3) None. Psychologist: "He gives us context to that, saying that she's the student body president and that she's the reason that Fordham University has a no beer or no drinking policy now, and that she is well-read and seems to be the type of student that the dean wants more of at Fordham University." answer: 1) None. 2) None. 3) None. Psychologist: "So they meet up at the steps of the library and dean runs off or something in a cape at least thats what he depicted it in his fabricated story again, and he was pied again, I think." answer: 1) None. 2) "So they meet up at the steps of the library" 3) None. Psychologist: "By running the story, he said hes going to do some embellishments on it and the embellishment he added was making him more of a superhero by giving him a catchphrase like, I am not an animal, but he said it in Latin. This guy was wearing a cape." answer: 1) None. 2) None. 3) None.
Eyespy	Psychologist: "Ferry people started getting more and more close and like easy to see and it said there on became just like little jumping dots. They were I'm assuming screaming at them for to come back." answer: 1) None. 2) "screaming at them for to come back." 3) None. Psychologist: "there was a specific event where they would go on vacation on I think she said Nantucket Island I forgot where it was" answer: 1) "Nantucket Island". 2) None. 3) None. Psychologist: "they were telling the story to those in the family. As a punishment to the kids, the kids were not allowed to play in the water. They had to stand on the beach and the kids were like, 'oh like can we just go in the water?' And they were like 'fine you can go'." answer: 1) None. 2) None. 3) "kids were like, 'oh like can we just go in the water?' And they were like 'fine you can go.'" Psychologist: "their whole her whole family was quite obsessed with the Kennedys, to the point that when JFK died she and her sister were in the car on a road trip of some sort." answer: 1) "when JFK died". 2) None. 3) None. Psychologist: "Then, shes talking about their summers where they would go to the beach with their aunts and cousins and uncles, and at the beach, the aunts and uncles loved to watch not the beach but the Kennedy house because the beach house that they would stay at was directly across from the Kennedys beach house." answer: 1) None. 2) None. 3) None. Psychologist: "I didn't know the last video guys name, but her name was Mikayla and she knocked out her uncles - I think her uncle most likely, his eyeball. And they then the eyeball. They then got punished, whatever." answer: 1) None. 2) "They then got punished" 3) None. Psychologist: "Then, the girl and her brother or cousins went out into the water, and they were messing around too much, and they got on a raft and the riptide pulled them under." answer: 1) None. 2) None. 3) "and they were messing around too much" Psychologist: "The aunts get mad and they were like, Okay, you guys are not allowed to touch the water in the beach, just the sand." answer: 1) None. 2) None. 3) None.
Oregontrail	Psychologist: "She was a teacher and then she became a comedian. Her kids apparently really liked the Oregon Trail and they liked the thought of dying." answer: 1) None. 2) "and then she became a comedian." 3) None. Psychologist: "It ended up being a baby wolf named, Benny-the-something, Benny-the-Chinny or something like that." answer: 1) None. 2) None. 3) "Benny-the-Chinny or something like that" Psychologist: "She compared - she did some parallels between Oregon Trail and the Dragons, what is it called, Dungeon and Dragons game. Every day they were talking about different things. The kids loved how dangerous." answer: 1) None. 2) None. 3) None. Psychologist: "When the kid rolled that she was going to die, another kid remembered, Hey, Im a doctor! So the teacher came over and was like, Its great! Come up here and help this child.So, they rolled and I think they rolled the wrong number." answer: 1) "they rolled the wrong number" 2) None. 3) "so the teacher came over and was like, Its great!" Psychologist: "I believe they ended up with a three and theyve obtained a wolf. Lets see." answer: 1) None. 2) "they ended up with a three" 3) None. Psychologist: "The teacher responded by saying that the wagon was, like I said, dedicated or reserved for carrying either a wounded, the sick, or supplies and that children never rode the cart." answer: 1) None. 2) None. 3) None. Psychologist: "They had like 790 miles to go to get to Oregon City, and she said they made it incredibly fast." answer: 1) "790 miles" 2) None. 3) None.

Continued on next page

	In-context learning examples
	Psychologist: "They would play different games where they would roll dice and after each time they roll the dice, it would be a different scenario and they had to make different choices." answer: 1) None. 2) None. 3) "after each time they roll the dice, it would be a different scenario and they had to make different choices"
Baseball	Psychologist: "Joe Varley was standing and leaning, looking pensively out a window as he writes a letter he just got in the post." answer: 1) "Joe Varley". 2) "looking pensively out a window as he writes a letter" 3) None. Psychologist: "I think I remember him talking about earning more money if he were to be a part of the Pittston's team earning more money and fame." answer: 1) "the Pittston's team" 2) None. 3) None. Psychologist: "He's going to find out he's a pitcher, and as he's teasing his sister and swinging around his arm like he's pitching, their mother walks into the room, she's about to water the plants, he knocks the water pot out of her hands, gets her involved." answer: 1) None. 2) None. 3) None. Psychologist: "So, the sister doesn't really want to go because of that, but he convinces her to go with him anyway to go meet them." answer: 1) "but he convinces her to go with him anyway to go meet them." 2) None. 3) None. Psychologist: "Oh yes, it was also revealed to be a snowy day and I believe the sister, Clara, mentioned something about the family or the father having debt, presumably a deceased father, something about a business." answer: 1) None. 2) "the father having debt, presumably a deceased father" 3) None. Psychologist: "He just said, 'I guess I'll get paid more,' and then I think he finally told her that he was going to get paid more and she got excited." answer: 1) None. 2) None. 3) None. Psychologist: "His mother says that that she doesn't quite like that idea or something along those lines, basically, saying that she refused the topic or the legitimacy of it." answer: 1) None. 2) None. 3) "saying that she refused the topic or the legitimacy of it." Psychologist: "He comes back in X, Y, Z hour, and he goes like, 'Hey,' and he goes, 'Joe, the train is not going to get in this time around.' He was like, 'Oh, what do you mean? Has it been derailed?'" answer: 1) None. 2) None. 3) "Has it been derailed?"

Table S3. In-context learning examples for each story. These examples are concatenated to the instruction and story transcript in Table S2, as part of the system prompt in the memory type classification model.

	Prompt content
Instruction	You are a helpful agent. Here is a story splitted into a list of events in order. For each user message, your job is to identify which event the user is talking about. Please respond with the event index.
Segmented story text	1. I began my illustrious career in journalism in the Bronx where I toiled as a hard-boiled reporter for the Ram, the student newspaper at Fordham University. 2. And one day I'm walking toward the campus center and out comes the elusive Dean McGowan, architect of a policy to replace Fordham's traditionally working- to middle-class students with wealthier, more prestigious ones. ...

Table S4. System prompt template for the event matching model. The system prompt consists of two sections: instruction and the segmented story transcript. The story transcript varies based on the story; we include the first two events from *Pieman* for reference.

	Beta	CI (Wald)	p-value
Recall: 229 participants (16,345 trials)			
Intercept	0.08	–0.31, 0.48	0.681
Centrality	0.06	0.01, 0.10	**0.011**
Contextual surprisal	–0.03	–0.08, 0.01	0.146
Sim. to prior corpus	0.06	0.01, 0.10	**0.013**
Event length	0.30	0.26, 0.34	**<0.001**
Serial position	–0.02	–0.06, 0.01	0.177
Factual error: 229 participants (8,295 trials)			
Intercept	–1.21	–1.48, –0.94	**<0.001**
Centrality	–0.14	–0.21, –0.06	**<0.001**
Contextual surprisal	0.02	–0.06, 0.9	0.661
Sim. to prior corpus	0.15	0.08, 0.23	**<0.001**
Event length	0.13	0.07, 0.18	**<0.001**
Serial position	–0.17	–0.23, –0.12	**<0.001**
Confabulation: 229 participants (8,295 trials)			
Intercept	–2.63	–2.85, –2.41	**<0.001**
Centrality	0.04	–0.07, 0.15	0.494
Contextual surprisal	0.16	0.06, 0.26	**0.003**
Sim. to prior corpus	0.00	–0.12, 0.11	0.953
Event length	0.14	0.05, 0.22	**0.002**
Serial position	0.08	0.00, 0.17	0.059

Table S5. Linear mixed-effects model results predicting recall and false memory outcomes using semantic predictors derived from the Universal Sentence Encoder and GPT-2 small. Standardized regression coefficients (β), 95% Wald confidence intervals (CIs), and p-values are reported for each fixed effect in the three models. Fixed effects include event-level semantic centrality, contextual surprisal, similarity to prior narrative corpus, event length, and serial position. Models include random intercepts for participants and stories. Significant predictors ($p < .05$) are shown in bold.

	Beta	CI (Wald)	p-value
Recall: 229 participants (16,345 trials)			
Intercept	0.08	–0.30, 0.46	0.672
Centrality	0.90	0.04, 0.13	**<0.001**
Contextual surprisal	–0.10	–0.15, –0.06	**<0.001**
Sim. to prior corpus	–0.06	–0.10, –0.02	**0.003**
Event length	0.32	0.28, 0.36	**<0.001**
Serial position	0.03	–0.01, 0.06	0.121
Factual error: 229 participants (8,295 trials)			
Intercept	–1.21	–1.49, –0.93	**<0.001**
Centrality	–0.13	–0.21, –0.06	**<0.001**
Contextual surprisal	0.03	–0.05, 0.11	0.456
Sim. to prior corpus	0.15	0.08, 0.23	**<0.001**
Event length	0.13	0.07, 0.18	**<0.001**
Serial position	–0.17	–0.23, –0.11	**<0.001**
Confabulation: 229 participants (8,295 trials)			
Intercept	–2.63	–2.84, –2.41	**<0.001**
Centrality	0.03	–0.09, 0.14	0.643
Contextual surprisal	0.12	0.01, 0.23	**0.028**
Sim. to prior corpus	0.00	–0.12, 0.11	0.937
Event length	0.15	0.06, 0.23	**0.001**
Serial position	0.06	–0.02, 0.15	0.130

Table S6. Linear mixed-effects model results predicting recall, factual error, and confabulation outcomes using semantic predictors derived from the Universal Sentence Encoder and Pythia. Standardized regression coefficients (β), 95% Wald confidence intervals (CIs), and p-values are reported for each fixed effect in the three models. Fixed effects include event-level semantic centrality, contextual surprisal, similarity to prior narrative corpus, event length, and serial position. Models include random intercepts for participants and stories. Significant predictors ($p < .05$) are shown in bold.

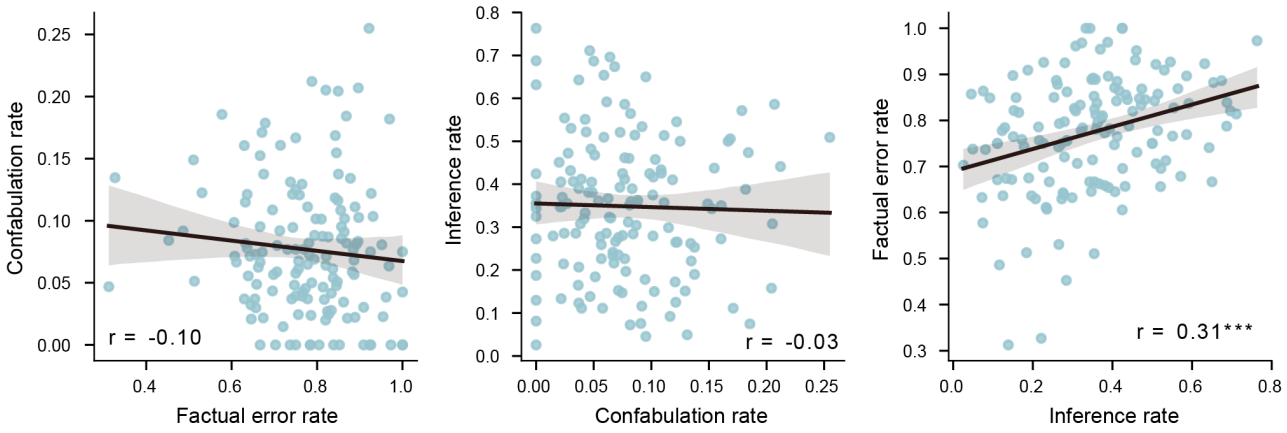

Figure S1. Correlations for predicted outcomes of narrative recall. Each point represents a story event; lines show linear fits with 95% confidence intervals. Confabulation and factual error rates were weakly negatively correlated (r = –0.10), inference and confabulation rates showed no relation (r = –0.03), and inference and factual error rates were positively correlated (r = 0.31, $p < .001$).

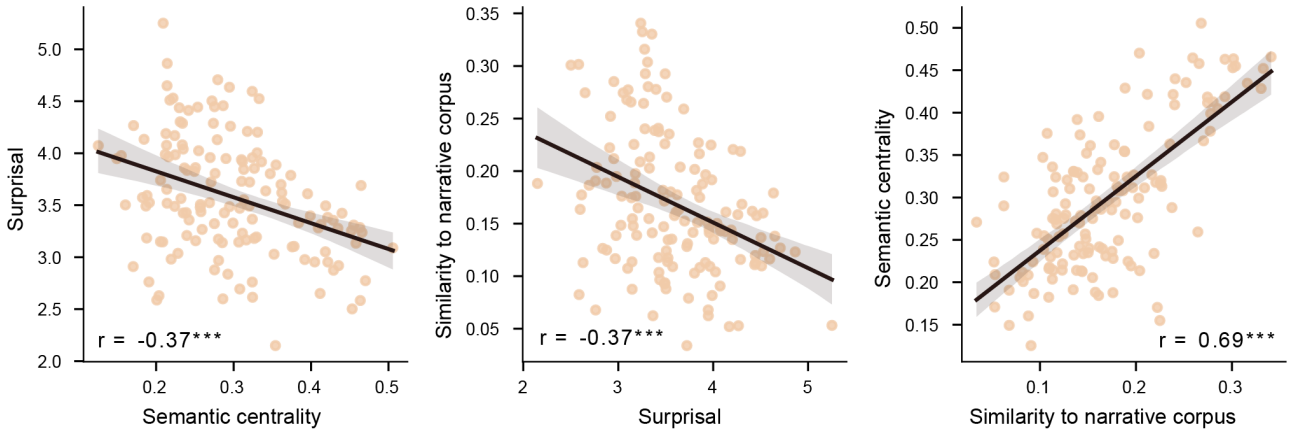

Figure S2. Correlations among semantic predictors. Each point represents a story unit; lines show linear fits with 95% confidence intervals. Semantic centrality was negatively correlated with surprisal (r = –0.37, $p < .001$) and positively correlated with similarity to narrative corpus (r = 0.69, $p < .001$). Surprisal and similarity were also negatively correlated (r = –0.37, $p < .001$).

References

1. Bartlett, F. A. A study in experimental and social psychology (1932).
2. Roediger, H. L. & McDermott, K. B. Creating false memories: Remembering words not presented in lists. *J. experimental psychology: Learn. Mem. Cogn.* **21**, 803 (1995).
3. Loftus, E. F. & Pickrell, J. E. The formation of false memories (1995).
4. Raccah, O., Chen, P., Gureckis, T. M., Poeppel, D. & Vo, V. A. The "Naturalistic Free Recall" dataset: Four stories, hundreds of participants, and high-fidelity transcriptions. *Sci. Data* **11**, 1317, 10.1038/s41597-024-04082-6 (2024).
5. Brainerd, C. J. & Reyna, V. F. *The science of false memory* (Oxford University Press, 2005).
6. Kimball, D. R. & Bjork, R. A. Influences of intentional and unintentional forgetting on false memories. *J. Exp. Psychol. Gen.* **131**, 116 (2002).
7. Schacter, D. L. *The seven sins of memory: How the mind forgets and remembers* (HMH, 2002).
8. Schacter, D. L. & Dodson, C. S. Misattribution, false recognition and the sins of memory. *Philos. Transactions Royal Soc. London. Ser. B: Biol. Sci.* **356**, 1385–1393 (2001).
9. Jacoby, L. L., Kelly, C. M. & Dywan, J. Memory attributions. In *Varieties of memory and consciousness*, 391–422 (Psychology Press, 2014).
10. Schacter, D. L., Guerin, S. A. & St. Jacques, P. L. Memory distortion: An adaptive perspective. *Trends cognitive sciences* **15**, 467–474, 10.1016/j.tics.2011.08.004 (2011).
11. Garry, M., Manning, C. G., Loftus, E. F. & Sherman, S. J. Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychon. bulletin & review* **3**, 208–214 (1996).
12. Loftus, E. F. & Hoffman, H. G. Misinformation and memory: the creation of new memories. *J. experimental psychology: Gen.* **118**, 100 (1989).
13. Goff, L. M. & Roediger, H. L. Imagination inflation for action events: Repeated imaginings lead to illusory recollections. *Mem. & Cogn.* **26**, 20–33 (1998).
14. Reyna, V. F. A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgm. & Decis. Mak.* **7**, 332–359 (2012).
15. Deese, J. On the prediction of occurrence of particular verbal intrusions in immediate recall. *J. experimental psychology* **58**, 17 (1959).
16. Gallo, D. A. False memories and fantastic beliefs: 15 years of the drm illusion. *Mem. & cognition* **38**, 833–848 (2010).
17. Pezdek, K. & Lam, S. What research paradigms have cognitive psychologists used to study "false memory," and what are the implications of these choices? *Conscious. cognition* **16**, 2–17 (2007).
18. Wade, K. A. *et al.* False claims about false memory research. *Conscious. cognition* **16**, 18–28 (2007).
19. Pardilla-Delgado, E. & Payne, J. D. The deese-roediger-mcdermott (drm) task: A simple cognitive paradigm to investigate false memories in the laboratory. *J. visualized experiments: JoVE* 54793 (2017).
20. Lee, H. & Chen, J. Predicting memory from the network structure of naturalistic events. *Nat. Commun.* **13**, 4235 (2022).
21. Heusser, A. C., Fitzpatrick, P. C. & Manning, J. R. Geometric models reveal behavioural and neural signatures of transforming experiences into memories. *Nat. Hum. Behav.* **5**, 905–919, 10.1038/s41562-021-01051-6 (2021).
22. Silva, M., Baldassano, C. & Fuentemilla, L. Rapid memory reactivation at movie event boundaries promotes episodic encoding. *J. Neurosci.* **39**, 8538–8548 (2019).
23. Lee, H., Chen, J. & Hasson, U. A functional neuroimaging dataset acquired during naturalistic movie watching and narrated recall of a series of short cinematic films. *Data Brief* **46**, 108788 (2023).
24. Mihalcea, R. *et al.* How developments in natural language processing help us in understanding human behaviour. *Nat. Hum. Behav.* **8**, 1877–1889, 10.1038/s41562-024-01938-0 (2024).
25. Georgiou, A., Can, T., Katkov, M. & Tsodyks, M. Large-scale study of human memory for meaningful narratives. *Learn. & Mem.* **32**, a054043 (2025).
26. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* **118**, e2105646118 (2021).
27. Michelmann, S., Kumar, M., Norman, K. A. & Toneva, M. Large language models can segment narrative events similarly to humans. *Behav. Res. Methods* **57**, 1–13 (2025).
28. Panela, R. A., Barnett, A. J., Barense, M. D. & Herrmann, B. Event segmentation applications in large language model enabled automated recall assessments. *arXiv preprint arXiv:2502.13349* (2025).
29. Kahana, M. J. *Foundations of human memory* (OUP USA, 2012).

30. Fenerci, C., Cheng, Z., Addis, D. R., Bellana, B. & Sheldon, S. Studying memory narratives with natural language processing. *Trends Cogn. Sci.* (2025).

31. OpenAI. Gpt-4o. https://openai.com/research/gpt-4 (2024). Large language model.

32. Schacter, D. L. Adaptive constructive processes and the future of memory. *Am. Psychol.* **67**, 603 (2012).

33. Schacter, D. L. & Addis, D. R. The ghosts of past and future. *Nature* **445**, 27–27 (2007).

34. Carpenter, A. C. & Schacter, D. L. Flexible retrieval: When true inferences produce false memories. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 335 (2017).

35. Yu, Y., Qu, W.-Y., Li, N. & Guo, Z. Open-category classification by adversarial sample generation. *arXiv preprint arXiv:1705.08722* (2017).

36. Barcina-Blanco, M., Lobo, J. L., Garcia-Bringas, P. & Del Ser, J. Managing the unknown in machine learning: Definitions, related areas, recent advances, and prospects. *Neurocomputing* 128073 (2024).

37. Reyna, V. F., Corbin, J. C., Weldon, R. B. & Brainerd, C. J. How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *J. Appl. Res. Mem. Cogn.* **5**, 1–9, 10.1016/j.jarmac.2015.12.003 (2016).

38. Reyna, V. F. & Brainerd, C. J. Fuzzy-trace theory: An interim synthesis. *Learn. individual Differ.* **7**, 1–75 (1995).

39. Demszky, D. *et al.* Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701, 10.1038/s44159-023-00241-5 (2023).

40. Klus, J. *et al.* Modeling memories, predicting prospections: Automated scoring of autobiographical detail narration using large language models. .

41. Shen, J., Mire, J., Park, H. W., Breazeal, C. & Sap, M. Heart-felt narratives: Tracing empathy and narrative style in personal stories with llms. *arXiv preprint arXiv:2405.17633* (2024).

42. Coane, J. H., McBride, D. M., Termonen, M.-L. & Cutting, J. C. Categorical and associative relations increase false memory relative to purely associative relations. *Mem. & cognition* **44**, 37–49 (2016).

43. Yeung, R. C., Stastna, M. & Fernandes, M. A. Understanding autobiographical memory content using computational text analysis. *Memory* **30**, 1267–1287 (2022).

44. Sheldon, S. *et al.* Differences in the content and coherence of autobiographical memories between younger and older adults: Insights from text analysis. *Psychol. aging* **39**, 59 (2024).

45. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. *Adv. neural information processing systems* **33**, 16857–16867 (2020).

46. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. & Specia, L. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 1 (Association for Computational Linguistics, 2017).

47. OpenAI. Gpt-2 small. https://huggingface.co/openai-community/gpt2 (2019). Pretrained 117M parameter version of GPT-2.

48. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. 24.

49. Moskvichev, A. & Mai, K.-V. Narrativexl: a large-scale dataset for long-term memory models. *arXiv preprint arXiv:2305.13877* (2023).

50. Cann, D. R., McRae, K. & Katz, A. N. False recall in the deese–roediger–mcdermott paradigm: The roles of gist and associative strength. *Q. J. Exp. Psychol.* **64**, 1515–1542 (2011).

51. Montefinese, M., Zannino, G. D. & Ambrosini, E. Semantic similarity between old and new items produces false alarms in recognition memory. *Psychol. research* **79**, 785–794 (2015).

52. Coane, J. H. *et al.* Manipulations of list type in the drm paradigm: A review of how structural and conceptual similarity affect false memory. *Front. Psychol.* **12**, 668550 (2021).

53. Michelmann, S. *et al.* Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nat. Commun.* **12**, 5394 (2021).

54. Levy, R. Memory and surprisal in human sentence comprehension. In *Sentence processing*, 78–114 (Psychology Press, 2013).

55. Sinclair, A. H., Manalili, G. M., Brunec, I. K., Adcock, R. A. & Barense, M. D. Prediction Errors Disrupt Hippocampal Representations and Update Episodic Memories. *bioRxiv* 2020.09.29.319418, 10.1101/2020.09.29.319418 (2021).

56. Howard, M. W. & Kahana, M. J. Contextual variability and serial position effects in free recall. *J. Exp. Psychol. Learn. Mem. Cogn.* **25**, 923 (1999).

57. Cer, D. *et al.* Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

58. Biderman, S. *et al.* Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373* (2023).

59. Brainerd, C. J. & Reyna, V. F. Fuzzy-trace theory and false memory. *Curr. Dir. Psychol. Sci.* **11**, 164–169 (2002).

60. Schacter, D. L. The seven sins of memory: insights from psychology and cognitive neuroscience. *Am. psychologist* **54**, 182 (1999).

61. Movva, R., Koh, P. W. & Pierson, E. Annotation alignment: Comparing llm and human annotations of conversational safety. *arXiv preprint arXiv:2406.06369* (2024).

62. Zacks, J. M. & Tversky, B. Event structure in perception and conception. *Psychol. bulletin* **127**, 3 (2001).

63. Baldassano, C. *et al.* Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721 (2017).

64. Baldassano, C., Hasson, U. & Norman, K. A. Representation of real-world event schemas during narrative perception. *J. Neurosci.* **38**, 9689–9699 (2018).

65. Chen, J. & Bornstein, A. M. The causal structure and computational value of narratives. *Trends Cogn. Sci.* (2024).

66. Nie, A. *et al.* Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Adv. Neural Inf. Process. Syst.* **36**, 78360–78393 (2023).

67. Brainerd, C. J. & Reyna, V. F. Fuzzy-trace theory and children's false memories. *J. experimental child psychology* **71**, 81–129 (1998).

68. Carneiro, P., Garcia-Marques, L., Fernandez, A. & Albuquerque, P. Both associative activation and thematic extraction count, but thematic false memories are more easily rejected. *Memory* **22**, 1024–1040 (2014).

69. Talarico, J. M. & Rubin, D. C. Confidence, not consistency, characterizes flashbulb memories. *Psychol. science* **14**, 455–461 (2003).

70. Hirst, W. & Phelps, E. A. Flashbulb memories. *Curr. Dir. Psychol. Sci.* **25**, 36–41 (2016).

71. Roediger, H. L., Watson, J. M., McDermott, K. B. & Gallo, D. A. Factors that determine false recall: A multiple regression analysis. *Psychon. bulletin & review* **8**, 385–407 (2001).

72. Sergi, I., Senese, V. P., Pisani, M. & Nigro, G. Assessing activation of true and false memory traces: A study using the drm paradigm. *Psychol. Belg.* **54**, 171–179 (2014).

73. Schacter, D. L. The seven sins of memory: An update. *Memory* **30**, 37–42, 10.1080/09658211.2021.1873391 (2022).

74. Gatti, D., Rinaldi, L., Mazzoni, G. & Vecchi, T. Semantic and episodic processes differently predict false memories in the drm task. *Sci. Reports* **14**, 256 (2024).

75. Loftus, E. F. Reconstructing memory: The incredible eyewitness. *Jurimetrics J.* **15**, 188–193 (1975).

76. Loftus, E. F. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learn. & memory* **12**, 361–366 (2005).

77. DePrince, A. P., Allard, C. B., Oh, H. & Freyd, J. J. What's in a name for memory errors? implications and ethical issues arising from the use of the term" false memory" for errors in memory for details. *Ethics & Behav.* **14**, 201–233 (2004).

78. Otgaar, H., Howe, M. L. & Patihis, L. What science tells us about false and repressed memories. *Memory* **30**, 16–21 (2022).

79. Lin, S., Hilton, J. & Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

80. Liang, P. *et al.* Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

81. Dong, Q. *et al.* A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).

82. Dong, Q. *et al.* A survey on in-context learning. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128, 10.18653/v1/2024.emnlp-main.64 (Association for Computational Linguistics, Miami, Florida, USA, 2024).

83. De Leeuw, J. R. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behav. research methods* **47**, 1–12 (2015).

84. Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

85. Wilcox, E. G., Gauthier, J., Hu, J., Qian, P. & Levy, R. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912* (2020).

86. Raccah, O., Chen, P., Willke, T. L., Poeppel, D. & Vo, V. A. Memory in humans and deep language models: Linking hypotheses for model augmentation. Paper at Memory in Real and Artificial Intelligence Workshop, NeurIPS (2022).

87. Kumar, M. *et al.* Bayesian surprise predicts human event segmentation in story listening. *Cogn. science* **47**, e13343 (2023).

88. Shain, C. Word frequency and predictability dissociate in naturalistic reading. *Open Mind* **8**, 177–201 (2024).

89. Cohn-Sheehy, B. I. *et al.* The hippocampus constructs narrative memories across distant events. *Curr. Biol.* **31**, 4935–4945 (2021).

90. Cohn-Sheehy, B. I. *et al.* Narratives bridge the divide between distant events in episodic memory. *Mem. & Cogn.* **50**, 478–494 (2022).

91. Shen, X., Houser, T., Smith, D. V. & Murty, V. P. Machine-learning as a validated tool to characterize individual differences in free recall of naturalistic events. *Psychon. Bull. & Rev.* **30**, 308–316 (2023).

92. Oh, B.-D., Yue, S. & Schuler, W. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. *arXiv preprint arXiv:2402.02255* (2024).

93. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32** (2019).

94. Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

95. Bates, D. lme4: Linear mixed-effects models using eigen and s4. *R package version* **1**, 1 (2016).