LLM-based scoring of narrative memories reveals that emotional arousal enhances central information at the expense of peripheral information

Xinyue Pan
Department of Psychology
University of Chicago
Chicago, USA
xpan02@uchicago.edu

Jadyn Park
Department of Psychology
University of Chicago
Chicago, USA
jadynpark@uchicago.edu

Yuan Chang Leong
Department of Psychology
University of Chicago
Chicago, USA
ycleong@uchicago.edu

***Abstract*—Memory is essential for well-being, identity, belief formation, and social cognition. While emotionally intense experiences are better remembered, it remains unclear whether emotions affect central (i.e., core elements) or peripheral (i.e., incidental or less critical) information similarly. Prior research has been limited by the lack of standardized and automated tools for assessing memory content at scale. This study utilizes large language models (LLMs) to quantify both emotional arousal in narrative stimuli and memory fidelity for central and peripheral information in verbal free recall of the narrative. We show that emotion enhances the recall of central information while diminishing recall of peripheral information. Our work introduces a reproducible and scalable framework that can be used in future studies to systematically examine how emotional intensity shapes narrative memory, with implications for psychological theory, educational assessment, and clinical evaluation.**

Keywords—episodic memory, emotion, free recall, Large Language Model (LLMs)

I. Introduction

Emotionally intense events often leave us with strikingly vivid memories, such as remembering where you were when you heard about the assassination of John F. Kennedy [1] or the attacks on 9/11 [2]. In daily life, emotionally charged memories can arise from dynamic and engaging experiences: a tense conversation, a close football match, or a suspenseful movie. These everyday emotional experiences shape how we encode, organize, and recall information that influences decision-making [3], learning [4], and well-being [5], [6]. In extreme cases, individuals who have experienced traumatic events involuntarily ruminate on those negative experiences, significantly affecting their quality of life [6]. A deeper understanding of how emotion affects different dimensions of memory can inform the design of educational practices, guide strategies to improve decision-making and enhance mental health interventions.

Past research has shown that emotional information is typically better remembered than neutral information [7], [8]. For instance, a surprise pop quiz in class is typically remembered better than the surrounding, uneventful lectures. Emotional events are thought to enhance memory by increasing arousal, a state of psychological and physiological activation that modulates attention and memory processes [9], [10], [11], [12]. However, it remains unclear whether arousal enhances memory uniformly or selectively, and in particular whether it benefits central (i.e., core elements) or peripheral (i.e., incidental or less critical) information similarly. One view, the attention narrowing hypothesis, proposes that arousal acts as a selective amplifier, enhancing memory for goal-relevant or salient elements, while suppressing less relevant background information [9], [13]. For example, eyewitnesses at a crime scene attend to and remember the weapon at the expense of other contextual details (e.g., the color of the shirt the perpetrator was wearing) [14], [15]. In contrast, other evidence suggests that arousal enhances attention broadly, improving the recall of both central and peripheral elements [16]. Resolving this debate requires methods that can capture the richness of lived experiences and reveal the fine-grained structure of memory.

One limitation of past work is that it has primarily relied on static cues, such as pictures and words, to elicit emotional responses and multiple-choice questions to assess recall [13], [17], [18]. However, static images poorly approximate the richness of real-world experiences, and multiple-choice formats limit insight into the content and structure of what individuals actually remember. In response to these limitations, an increasing number of studies have adopted dynamic audiovisual stimuli (e.g., films, narratives) to better simulate real-life scenarios and have used free recall tasks to capture more comprehensive memory representations [19].

In particular, audiovisual narratives capture the complexity, temporal structure, and multimodal features of everyday experiences, enabling researchers to investigate how memory unfolds in dynamic, ecologically valid contexts. Narratives and films require individuals to process both perceptual (i.e., color, sound) and conceptual information (i.e., causal relationships) [20]. This design has allowed researchers to study how people segment continuous experience into discrete events [20], [21], how narratives drive fluctuations in emotional experience [22], and how these fluctuations affect memory [10], [23].

Although naturalistic stimuli and free recall capture the richness of real-world memory, that same richness makes them difficult and resource-intensive to analyze at scale. Annotating recall responses, aligning them with stimulus content, distinguishing central from peripheral details, and evaluating

memory fidelity all require extensive manual effort and are influenced by subjective judgment [24]. These burdens limit the feasibility of large-scale studies and reduce reproducibility across laboratories.

The present study addresses these challenges by introducing a generative artificial intelligence pipeline that quantifies emotional arousal in narrative stimuli and scores recall transcripts for memory fidelity across central and peripheral details. This approach combines the ecological validity of dynamic audiovisual stimuli with the scalability and reproducibility of automated computational methods. Applying this framework, we investigate whether emotional arousal enhances both narratively central and peripheral information. Beyond the present study, this pipeline offers a scalable framework that can be applied in domains such as education, clinical psychology, and forensic science, where large-scale and fine-grained analysis of memory content is critical.

II. Related Work

A. Emotion and Memory

Emotional events are generally better remembered than neutral ones [1], [2], [7], [10], [16]. For example, one might vividly recall a heated argument with their parents but have no memory of an ordinary conversation. Stronger emotional arousal captures attention and promotes deeper encoding, which strengthens later recall [8], [25]. However, how attention is distributed during an emotional experience remains less clear. In other words, although prior work consistently shows that individuals prioritize emotional events in memory, it is still debated which specific aspects of these events receive the greatest attentional and mnemonic priority.

The attentional narrowing hypothesis proposes that arousal selectively prioritizes goal-relevant or salient elements while suppressing peripheral, background, or associative details [9], [11], [13], [17]. One classic example is when eyewitnesses vividly recall the core feature of a threat at a crime scene (e.g., a gun), but poorly recall peripheral elements (e.g., the clothing of the perpetrator), an effect known as the "weapon focus" effect [14], [15], [17], [26]. In line with this hypothesis, arousal has been shown to enhance memory for an object but impairs memory for the spatial and temporal context in which it was presented, leading participants to recognize objects from emotionally arousing events yet misattribute where or when they appeared [27]. These studies suggest that emotional arousal captures attention, enhancing encoding of central information while disrupting memory of peripheral or contextual features.

However, not all evidence supports the view that arousal narrows attention towards the central information. A two-week experiment employing an incidental learning procedure found that emotional content enhanced retention of both plot-relevant and plot-irrelevant details, challenging the assumption that emotional arousal impairs memory for peripheral information [16]. Such findings motivate an alternative view, especially under extended experience, that emotional arousal may function less like a narrowing spotlight and more like a global amplifier that strengthens encoding across the entire experience.

First, emotional arousal may strengthen the process by which different elements of an experience are bound together into a coherent memory trace. When people encode a story, central and peripheral details may be integrated through their spatial or temporal association [28]. In other words, peripheral contextual details may get to "ride along" with the core storyline as part of an integrated memory representation [12], [29]. Second, many peripheral details (e.g., settings, props, or background character behaviors) may fit naturally into familiar schemas. Schema-congruent information is typically well-remembered [30]. Thus, enhanced memory of an event under heightened arousal may similarly improve the incidental details that are consistent with existing knowledge structures.

Few studies have tested these possibilities with naturalistic stimuli that better capture the causal structure, semantic richness, and temporal unfolding of everyday emotional experiences. Thus, it raises the critical question of *whether arousal in extended narratives selectively strengthens memory for central elements at the expense of peripheral ones or instead amplifies memory broadly*. To address this question, the present study examines the free recall of emotionally rich video clips.

B. LLMs Usage in Free Recall Analysis

Large Language Models (LLMs) are increasingly validated as computational assistants in psychological and memory research. These models capture patterns in language that reflect higher-level human cognition and behavior, including emotions, ideology, and moral judgment [31], [32]. For example, generative pre-trained (GPT) models may grasp broader contextual understandings in language use, which allows the model to outperform dictionary-based methods in labeling emotions, sentiments, offensiveness, and moral judgment [33].

Beyond word- or sentence-level semantics, LLMs also exhibit sensitivity to narrative structures. Narratives are not merely strings of words but organized sequences of events, and a key cognitive process in comprehending them is event segmentation-identifying meaningful boundaries in an unfolding story, such as shifts in scene or action [20]. Segmentations generated by GPT have been found to be consistent with those produced by human raters [33]. This suggests that LLMs recognize not only the emotional and semantic content of language but also higher-order narrative structure. The ability to parse narrative structure suggests that LLMs may also be useful for distinguishing between central details that drive the story forward versus peripheral content.

LLMs are not only cognitively aligned with human judgments but also offer methodological advantages. Prior work has demonstrated that Open-AI's GPT models achieve high levels of accuracy in complex language tasks without requiring extensive task-specific pre-training, suggesting that their representations capture broadly generalizable features of human cognition [33]. In addition, LLM outputs are both consistent and scalable. This addresses longstanding challenges associated with human annotation, including

variability across raters, the high cost of training coders, and the practical limits on the volume of data that can be processed. By contrast, LLMs can generate stable annotations across very large datasets, enabling reproducible analyses that would be infeasible with human-only approaches.

Building on these capacities, the present study leverages LLMs to quantify both the emotional arousal elicited by narrative events and participants' memory fidelity for those events. First, LLMs were used to estimate emotional arousal in the stimuli, providing a consistent and scalable alternative to subjective ratings. Second, we distinguished narratively central, plot-driving elements from peripheral descriptive details and compared them against free recall transcripts, which allowed the model to score each recalled element systematically. This framework enables cost-efficient and reproducible analyses of how emotional arousal shapes memory in ecologically valid, naturalistic contexts.

III. Data and Methods

This project investigates how emotional arousal influences the memory fidelity of central and peripheral details using open-ended recall data from existing studies that employed naturalistic narrative stimuli. We leverage large language models (LLMs) to systematically (1) rate the emotional arousal level of stimuli, and (2) score the fidelity of recall content.

A. Data Source

We use two open-access datasets that include narrative stimuli and free verbal recall: the Sherlock and Film Festival datasets. These data also include concurrent functional magnetic resonance imaging recordings, which were not analyzed in the present study. Both datasets include narrative event segmentations with accompanying written descriptions. Event-level emotional arousal scores were rated by human coders.

The Sherlock recall dataset includes verbal recall transcripts from 17 participants who watched a 50-minute episode of the BBC series Sherlock [34]. Participants were instructed to watch the episode as they normally would and were informed that they would later describe what they had seen [35]. Immediately after viewing, they verbally recounted the episode in as much detail as possible, aiming to follow the original order but prioritizing completeness over chronology. They were encouraged to speak for at least 10 minutes and allowed to continue without time limits. An independent annotator segmented the episode into 48 events based on major narrative changes (e.g., time, location, topic, characters) and provided written descriptions for each [35].

The Film Festival dataset includes verbal recall data from 15 participants who watched 10 short films (2.15–7.75 minutes each) varying in content and structure [36]. Participants were told to attend naturally and informed that they would later recall the films. Following viewing, they freely described the films in any order and as thoroughly as possible, with a recommended minimum of 10 minutes. For this study, the stimuli were re-segmented into 68 narrative events based on shifts in time, location, topic, or characters. These events were annotated with written descriptions, following the same procedure used for the Sherlock dataset. The datasets included transcripts of the verbal recalls that had been segmented into utterances and matched to events.

To increase generalizability and statistical power, we pooled data across the two datasets. Both datasets feature extended narrative stimuli with rich event structures, free verbal recall protocols, and independent event segmentations with written annotations. At the same time, the datasets differ in important ways. While Sherlock offers a single, long-form episode with continuous plot development, Film Festival contains multiple shorter film clips with diverse content and narrative styles. Pooling across the two datasets allowed us to test whether the effects of emotional arousal on memory generalize across both extended and varied narrative contexts, rather than being specific to a single stimulus or genre. In all subsequent analyses, the dataset was included as a random effect in mixed effects models to account for baseline differences between the two sources.

B. Constructing Measures of Emotional Arousal

We adapted a previous approach for rating the arousal of narrative events from text annotations [10]. Specifically, we prompted GPT-4o with a definition of arousal and tasked it to rate the arousal level of each narrative event:

> *Arousal refers to when you are feeling very mentally or physically alert, activated, and/or energized.*
>
> *Read the following description of a scene and rate the arousal level of the scene on a scale of 1 to 10, with 1 being low arousal and 10 being high arousal.*
>
> *Please give a numeric rating. Only give the rating; no need to provide explanations.*
>
> *Scene: {annotation}*

GPT-4o was run with a temperature of 0.0, producing deterministic scores for each event. To validate these ratings, we relied on a set of human arousal ratings of these events collected from 30 participants. The 30 ratings were z-scored within participants and averaged to obtain a single arousal rating for each event. We then computed the correlation between the GPT-4o ratings and the human ratings.

C. Scoring Memory Fidelity

Our central construct of interest was memory fidelity, defined as the degree to which participants accurately recalled specific narrative details from the original event. From prior work, we know that emotional memories may feel subjectively vivid yet often diverge from the original experience in accuracy [7]. Fidelity, therefore, differs from broader measures such as recall volume or semantic similarity because it focuses on whether particular details are faithfully reproduced in memory rather than how much is remembered.

We assessed the recall fidelity for both central and peripheral details. Central details were defined as causally essential elements of a narrative that sustain the storyline, such as events that drive the plot forward, explain character motivations, or mark turning points. Peripheral details were defined as descriptive elements that enrich the narrative

context (e.g., setting, atmosphere, incidental features) but are not required for the causal progression of the story.

As illustrated in Fig. 1, each movie scene was broken down into a set of small, discrete 'detail elements.' These include central elements that drive the plot forward and peripheral elements that add descriptive context. Each element was treated as a separate unit (e.g., C1, C2 for central; P1, P2 for peripheral), creating a checklist against which recall transcripts could be evaluated.

We implemented this as a two-stage scoring procedure using GPT-4o. In the first stage, the model was prompted to produce lists of central and peripheral elements. The prompt included explicit definitions of central and peripheral details:

> *Central details are causally essential elements of a narrative that sustain the storyline. They include information that drives the plot forward, explains character motivations, or marks turning points in the story. Without these details, the coherence or progression of the narrative would be disrupted.*
>
> *Peripheral details are descriptive elements that enrich the narrative context but are not essential to its causal structure. They provide texture, atmosphere, or background information (e.g., setting descriptions or incidental features), yet their absence would not alter the core storyline or change character motivations.*

The model was also provided with a summary of the overall narrative, the text annotation of the event to be scored, and instructions to output non-redundant units with less than 10 words each. The number of central and peripheral detail elements was held constant for each event but allowed to vary across events.

Each participant's free recall transcript was then evaluated against the generated checklists. GPT-4o was instructed to act as an expert annotator and was provided with both the transcript and the corresponding list of central or peripheral detail elements for a given event. For each detail, the model assigned a score on a 3-point scale (2 = present, 1 = partially present, 0 = absent). Scores were returned in a structured Markdown table, with one row per detail and columns for *participant ID*, *event number*, *detail ID*, and *score*. This standardized output ensured consistency across participants and facilitated seamless integration into the analysis pipeline.

Full prompt templates for both stages are provided in the Appendix. The appendix illustrates the general structure of the prompts, while the full set of templates and completions for all narrative events is available in the anonymized GitHub repository[1].

Scores were then averaged across available elements within each event, separately for central and peripheral categories, yielding event-level fidelity metrics. This design ensured that variability reflected the number of items recalled rather than verbosity. This produces interpretable, element-level scores that parallel the logic of multiple-choice testing, without introducing the cueing effects of explicit response options.

For comparison, we also computed a text similarity measure commonly used in prior work (e.g., [10]), which estimated overall recall quality as the cosine similarity between participant transcripts and event annotations encoded with Google's Universal Sentence Encoder (USE) [10]. Although embedding-based similarity offers an index of overall recall quality, it does not distinguish between central and peripheral details. Here, we tested whether our central and peripheral fidelity scores independently predicted variation in USE-based similarity using a linear mixed effects model.

D. Arousal Effects on Memory Fidelity

Recall scores for central and peripheral elements were then analyzed separately to test how emotional arousal influenced the memory encoding of central and peripheral detail elements. To do this, we fit Bayesian mixed-effect models to ask whether event-level emotional arousal (rated by GPT-4o on a 1-10 scale) predicts differences in recall scores (how much of each event's details were remembered on a 0-2 scale). The analysis was run separately for central and peripheral details, and with data pooled across the Sherlock recall and Film Festival datasets to increase statistical power. We included random intercepts for participants and stimuli to account for random variation across participants and stimuli, as well as the hierarchical structure of the data. We also replicated this analysis using human-rated emotional arousal.

Models were estimated in *R* version 4.4.1 using the *brms* package (version 2.22.0), with random seed "123" to ensure reproducibility and weakly informative priors to facilitate convergence. Posterior distributions were obtained via four Markov Chain Monte Carlo (MCMC) chains, each with 5,000 iterations, of which the first 2,000 iterations were discarded as burn-in. We checked for model convergence by ensuring that the Gelman-Rubin diagnostic $\hat{R}$ was less than 1.1 for all parameters. For each parameter of interest, we report the posterior mean, standard deviation, and 95% credible interval (CrI), alongside posterior plots visualizing the uncertainty in the arousal effect.

To evaluate model fit, we compared the Widely Applicable Information Criterion (WAIC) of our models against null models excluding the predictors of interest, using the *loo* package (version 2.8.0). Lower WAIC values indicate better out-of-sample predictive performance. In addition, we computed Bayes Factors (BF_{10}) with the *bayestestR* package (version 0.16.1) to assess the strength of evidence relative to the null model. A BF_{10} greater than 1 indicates support for the alternative model, with increasing values corresponding to stronger evidence. Full model formulas and prior specifications are accessible in the Appendix.

[1] GitHub repository link for review:
https://github.com/xpan4869/llm-memory-scoring

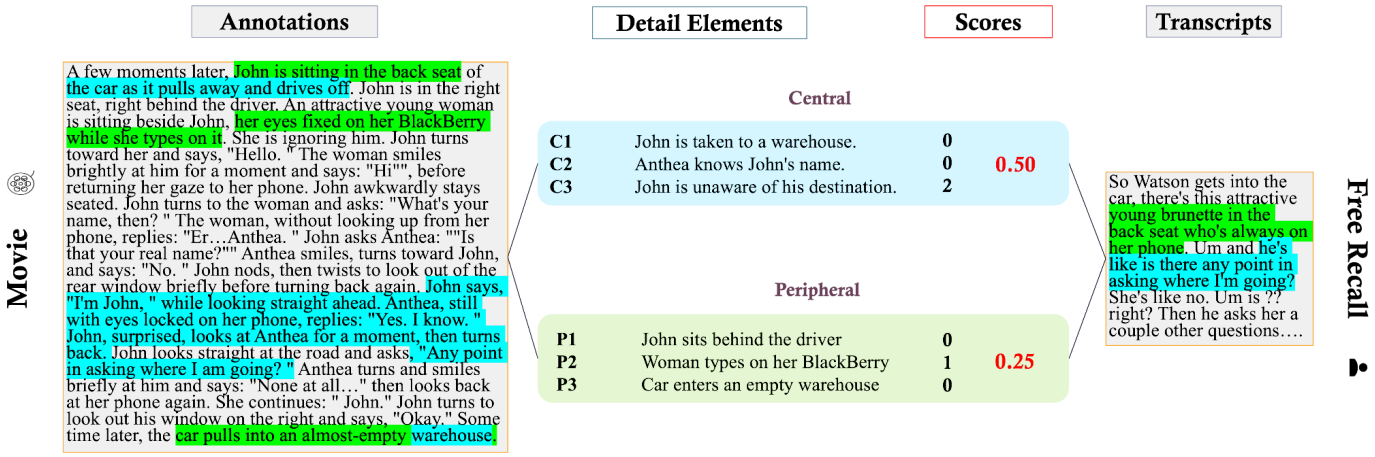

Fig. 1. Example of the annotation and scoring pipeline. This is a scene in the middle of the first episode of *Sherlock* where John Watson was led to a warehouse to meet with Mycroft Holmes. Narrative annotations (left) were decomposed into central and peripheral detail elements (middle). Each element was scored against participant free recall (right) on a 3-point scale (2 = present, 1 = partially present, 0 = absent). Scores were then averaged for each event to produce corresponding memory fidelity indices (red numbers) for central and peripheral details.

IV. Results

We combined the Sherlock and Film Festival datasets to increase statistical power. Both stimuli were segmented into a total of 116 narrative events defined by major shifts in the storylines (see Data and Methods).

Human behavioral arousal ratings showed strong consistency across participants (*Film Festival*: median = 3.2, range = 1.3–4.7, one-to-average r=.72, p<.01; *Sherlock*: median = 3, range = 1.5–4.7, one-to-average r=.69, p<.01), indicating shared perceptions of the emotional intensity of narrative events. Moreover, human behavioral arousal ratings (median = 3.10, range = 1.39–4.67) and LLM-generated ratings (median = 2.50, range = 1.00–4.50) were positively correlated (r = .74, p < .001), supporting the convergent validity of our LLM-based approach for obtaining emotional arousal ratings.

For each event, we prompted GPT-4o to generate a balanced set of central and peripheral detail elements, with the number of detail elements per category ranging from one to nine, depending on the event. Given these lists of central and peripheral detail elements, we then provided GPT-4o with each event's detail set and corresponding recall transcript to assign a score to each detail on a three-point scale (0, 1, or 2). For each event, we then averaged across its detail elements to obtain a central score and a peripheral details score. On average, participants recalled significantly more central details (M = 0.75, SD = 0.22) than peripheral details (M = 0.27, SD = 0.15, $t(31) = 28.8$, p < .001).

To assess whether central and peripheral details each make independent contributions to overall recall quality, we tested their relationship to a continuous measure of transcript–annotation overlap. Specifically, we computed cosine similarity between each recall transcript and its corresponding annotation by encoding both with Google's Universal Sentence Encoder (USE), following prior work [10].

In the same regression model including both predictors, central detail scores (β = 0.26, p < .001) and peripheral detail scores (β = 0.10, p < .001) each independently predicted cosine similarity. These results indicate that both central and peripheral recall contribute unique variance to overall recall quality, supporting the view that each dimension captures non-redundant aspects of memory fidelity.

We then examined whether LLM-rated emotional arousal predicts memory recall for central and peripheral details. A Bayesian multilevel model indicated that the recall of central information was higher for events that were rated as more emotionally arousing (b = 0.04, 95% HDI = [0.02, 0.07], p(b>0) = 1, WAIC = 3094, $WAIC_{null}$ = 3103, BF_{10} = 3.21; Fig. 2A).

In contrast, recall of peripheral information was lower for more emotionally arousing events (b = -0.04, 95% HDI = [-0.05, -0.02], p(b<0) = 1, WAIC = 1560, $WAIC_{null}$ = 1577, BF_{10} = 18.15; Fig. 2B). Taken together, these results suggest that emotional arousal selectively enhances memory for central information while impairing recall of peripheral details.

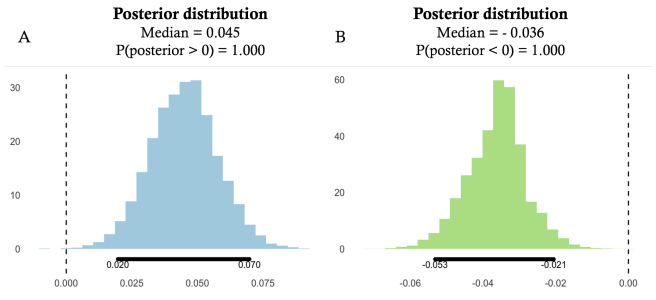

Fig. 2. Posterior distribution and model fits of the effect of LLM-rated arousal on memory fidelity. (A) Posterior distribution of the arousal effect on central recall; (B) Posterior distribution of the arousal effect on peripheral recall.

To assess the robustness of these findings, we then examined models using human-rated arousal scores. For central details, a Bayesian multilevel model indicated that arousal was positively associated with the number of details recalled, consistent with the LLM-based results (b = 0.11, 95% HDI = [0.08, 0.15], p(b>0) = 1, WAIC = 3067, $WAIC_{null}$ = 3104, $BF_{10} = 7.46 \times 10^3$).

For peripheral details, the results pointed toward a negative relationship between arousal and recall, again consistent with LLM-based results (b = -0.01, 95% HDI = [-0.04, 0.01], p(b<0) = 0.86, WAIC = 1578, $WAIC_{null}$ = 1577, BF_{10} = 0.021), but the evidence did not reach the pre-specified probability threshold of 0.95. We speculate that the human-based results may be less reliable as they depend on participants' metacognitive judgments and can vary subjectively across samples. Overall, these supplementary analyses indicate that results using the human-rated arousal broadly converge with the LLM-based results.

V. Discussion

While it has been consistently shown that emotional information is remembered more vividly and accurately [1], [2], [7], [10], [16], there has been debate on which memory features are better remembered [9], [11], [13], [16], [26], [27]. Our findings suggest that emotional arousal enhances memory fidelity for central information while suppressing peripheral information in naturalistic narratives. Rather than uniformly boosting all aspects of a story, higher emotional intensity appears to prioritize what is central to the plot while downplaying contextual details.

A few mechanisms may contribute to why arousal enhances central information in narratives. First, arousal has been shown to bias attention toward motivationally relevant cues [9], [26], [37], [38]. In the context of narratives, this may refer to the causally important elements of a story (i.e., who did what to whom and why). Second, central details are also more likely to be the source of arousal themselves. Plot twists, conflicts, and moments of threat are often narratively central as well as emotionally engaging. When arousal is elicited by these elements, it further amplifies their salience, ensuring they capture more attentional resources and are processed more deeply [9], [37]. Third, central details are typically more causally and thematically integrated within the narrative than peripheral ones. As they connect to multiple parts of the storyline, they can be retrieved through several associative pathways. Emotional arousal may strengthen these associative links, making it more likely that recalling one part of the story will reinstate related central details.

However, mental resources are limited. In biasing attention toward central details, arousal reduces the processing resources available for peripheral features, which are encoded less effectively as a result [9], [26]. Furthermore, peripheral information that lacks strong causal or semantic integration within a schema or narrative is more likely to be distorted [39], [40]. Finally, because peripheral features receive little reinforcement during consolidation, they are less likely to be stabilized in memory [41], [42], [43]. This diminishing effect reflects the cost of emotional prioritization: even though central details are better remembered, peripheral features become more susceptible to forgetting.

Together, the present study illustrates a tradeoff in memory between central and peripheral information under heightened emotional arousal, extending beyond individual variability and laboratory exposures.

Although individuals vary in their emotional responses, we observed strong correlations across participants and alignment between human and LLM-generated ratings. This suggests that shared narrative structures evoke reliable patterns of arousal, strengthening the validity of our results. Nevertheless, future work should continue to investigate how idiosyncratic factors shape the balance of central and peripheral recall. Beyond these considerations, caution is also warranted when generalizing our results to older populations, as our sample consisted primarily of young adults and age-related differences in memory are well-documented [5].

The tradeoff effect was evident when pooling from two distinct narrative contexts: a long-form, continuous television episode (*Sherlock*) and a set of shorter, stylistically diverse films (*Film Festival*). Combining these datasets allowed us to test whether arousal-driven memory effects generalize across variations in narrative length, genre, and style, rather than being tied to a single stimulus. The convergence of findings across both datasets supports the robustness of the observed tradeoff, though it should be noted that film-based narratives still reflect a specific cultural format and may not fully capture autobiographical or personally significant experiences. Extending current work to autobiographical contexts may therefore provide further insights into the mechanisms of memory encoding in everyday life.

Recognizing this tradeoff is essential when considering how emotional arousal can be leveraged to boost memory encoding, since information that feels peripheral in the moment may be essential in hindsight. For example, a student might vividly recall a dramatic classroom demonstration but fail to retain the underlying formula or principle it was meant to illustrate.

The tradeoff also resonates with clinical observations. Survivors of traumatic events often vividly recall the emotional core of the experience while failing to anchor it in the surrounding time and place [44], [45]. Such an imbalance reflects the same selective prioritization seen in our data: arousal strengthens memory for central, emotionally charged details while weakening recall of contextual information. Although our study does not directly address trauma, this parallel illustrates how the mechanisms observed in controlled narrative settings may scale to autobiographical memory in real life.

Beyond theoretical insights, our pipeline provides a reproducible and interpretable method for quantifying central and peripheral recall at scale, substantially reducing the time and cost of traditional, labor‑intensive annotation.

In education, assessing both central and peripheral recall goes beyond testing rote facts, offering a scalable way to gauge whether learners build richer, transferable understanding. In clinical practice, quantifying central versus peripheral recall may illuminate how disorders such as PTSD bias memory encoding and retrieval. In developmental and aging research, scalable annotation allows tracking how children, older adults, or clinical populations differentially weight core versus incidental information. Because memory fidelity shapes self-narrative and identity, this pipeline also opens avenues for studying cultural and individual differences in autobiographical memory construction.

To our knowledge, no existing model or publicly available human-coded dataset systematically differentiates between central and peripheral narrative elements. While our approach demonstrates that LLMs can reproduce and extend this distinction in free recall data, future work should incorporate independent human coding to establish a stronger benchmark. Such efforts would mitigate concerns about potential data contamination from publicly available annotations and provide a more rigorous test of the method's generalizability.

Another limitation of the present study is the modest sample size. Our analyses were based on a total of 32 participants, with 17 from the Sherlock dataset and 15 from the Film Festival dataset. While such numbers may appear insufficient to justify the application of large language models, it is important to recognize that even datasets at this scale demand extensive annotation for both emotional arousal and memory fidelity to enable systematic analysis. Because manual approaches are labor-intensive and prone to inconsistency, automated methods provide opportunities for standardized and reproducible pipelines in narrative memory research. Larger open-source datasets such as the *Four Stories* dataset (n=229) exist, but these currently lack detailed narrative annotations required for our modeling framework [46].

Future work can overcome this limitation by further automation, leveraging large language models to generate narrative annotations. This would allow researchers to take advantage of these larger datasets, providing a stronger test for the generalizability of our results and expanding the scope of analysis to examine other aspects of memory. For example, future work could examine the temporal dynamics, including whether central and peripheral details differ in their forgetting rates over time, and investigate potential modulators, such as stress or variations in narrative style. These directions would extend our framework and advance our understanding of how emotional arousal shapes memory across contexts.

In sum, this study suggests that emotional arousal in naturalistic, causally structured narratives produces a selective enhancement of memory fidelity, strengthening central while diminishing peripheral information. These findings extend beyond laboratory demonstrations of attentional narrowing, highlighting the importance of ecological validity in emotion-memory research. Just as importantly, our LLM-based scoring pipeline provides a scalable, interpretable, and practical tool for investigating memory that overcomes the limits of human coding and enables large-scale, reproducible analyses. By advancing both theory and method, this work opens new avenues for investigating how emotions shape memory in education, clinical practice, and everyday life.

References

[1] R. Brown and J. Kulik, "Flashbulb memories," *Cognition*, vol. 5, no. 1, pp. 73–99, Jan. 1977, doi: 10.1016/0010-0277(77)90018-X.

[2] J. M. Talarico and D. C. Rubin, "Confidence, Not Consistency, Characterizes Flashbulb Memories," *Psychol. Sci.*, vol. 14, no. 5, pp. 455–461, Sept. 2003, doi: 10.1111/1467-9280.02453.

[3] C. R. Madan, "Memory Can Define Individual Beliefs and Identity—and Shape Society," *Policy Insights Behav. Brain Sci.*, vol. 11, no. 1, pp. 102–109, Mar. 2024, doi: 10.1177/23727322231220258.

[4] N. Cowan, "Working Memory Underpins Cognitive Development, Learning, and Education," *Educ. Psychol. Rev.*, vol. 26, no. 2, pp. 197–223, June 2014, doi: 10.1007/s10648-013-9246-y.

[5] C. J. Dinius, C. E. Pocknell, M. P. Caffrey, and R. A. P. Roche, "Cognitive interventions for memory and psychological well-being in aging and dementias," *Front. Psychol.*, vol. 14, Feb. 2023, doi: 10.3389/fpsyg.2023.1070012.

[6] J. A. Bisby, N. Burgess, and C. R. Brewin, "Reduced Memory Coherence for Negative Events and Its Relationship to Posttraumatic Stress Disorder," *Curr. Dir. Psychol. Sci.*, vol. 29, no. 3, pp. 267–272, June 2020, doi: 10.1177/0963721420917691.

[7] M. Lewis, Ed., *Handbook of emotions*, 3rd ed. New York, NY, USA: Guilford Press, 2008.

[8] N. A. Murphy and D. M. Isaacowitz, "Preferences for emotional information in older and younger adults: A meta-analysis of memory and attention tasks," *Psychol. Aging*, vol. 23, no. 2, pp. 263–286, June 2008, doi: 10.1037/0882-7974.23.2.263.

[9] M. Mather and M. R. Sutherland, "Arousal-Biased Competition in Perception and Memory," *Perspect. Psychol. Sci.*, vol. 6, no. 2, pp. 114–133, Mar. 2011, doi: 10.1177/1745691611400234.

[10] J. S. Park, K. Gollapudi, J. Ke, M. Nau, I. Pappas, and Y. C. Leong, "Emotional arousal enhances narrative memories through functional integration of large-scale brain networks," Mar. 13, 2025, *bioRxiv*. doi: 10.1101/2025.03.13.643125.

[11] E. A. Kensinger, R. J. Garoff-Eaton, and D. L. Schacter, "Memory for specific visual details can be enhanced by negative arousing content," *J. Mem. Lang.*, vol. 54, no. 1, pp. 99–112, Jan. 2006, doi: 10.1016/j.jml.2005.05.005.

[12] E. A. Phelps, "Human emotion and memory: interactions of the amygdala and hippocampal complex," *Curr. Opin. Neurobiol.*, vol. 14, no. 2, pp. 198–202, Apr. 2004, doi: 10.1016/j.conb.2004.03.015.

[13] A. Burke, F. Heuer, and D. Reisberg, "Remembering emotional events," *Mem. Cognit.*, vol. 20, no. 3, pp. 277–290, May 1992, doi: 10.3758/BF03199665.

[14] E. F. Loftus, G. R. Loftus, and J. Messo, "Some facts about 'weapon focus,'" *Law Hum. Behav.*, vol. 11, no. 1, pp. 55–62, Mar. 1987, doi: 10.1007/BF01044839.

[15] N. M. Steblay, "A meta-analytic review of the weapon focus effect," *Law Hum. Behav.*, vol. 16, no. 4, pp. 413–424, Aug. 1992, doi: 10.1007/BF02352267.

[16] F. Heuer and D. Reisberg, "Vivid memories of emotional events: The accuracy of remembered minutiae," *Mem. Cognit.*, vol. 18, no. 5, pp. 496–506, Sept. 1990, doi: 10.3758/BF03198482.

[17] S.-Å. Christianson, E. F. Loftus, H. Hoffman, and G. R. Loftus, "Eye fixations and memory for emotional events," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 17, no. 4, pp. 693–701, July 1991, doi: 10.1037/0278-7393.17.4.693.

[18] E. A. Kensinger, B. Brierley, N. Medford, J. H. Growdon, and S. Corkin, "Effects of normal aging and Alzheimer's disease on emotional memory," *Emotion*, vol. 2, no. 2, pp. 118–134, June 2002, doi: 10.1037/1528-3542.2.2.118.

[19] H. Lee, B. Bellana, and J. Chen, "What can narratives tell us about the neural bases of human memory?," *Curr. Opin. Behav. Sci.*, vol. 32, pp. 111–119, Apr. 2020, doi: 10.1016/j.cobeha.2020.02.007.
[20] J. M. Zacks and K. M. Swallow, "Event Segmentation," *Curr. Dir. Psychol. Sci.*, vol. 16, no. 2, pp. 80–84, Apr. 2007, doi: 10.1111/j.1467-8721.2007.00480.x.
[21] C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, and K. A. Norman, "Discovering Event Structure in Continuous Narrative Perception and Memory," *Neuron*, vol. 95, no. 3, pp. 709-721.e5, Aug. 2017, doi: 10.1016/j.neuron.2017.06.041.
[22] J. Ke, H. Song, Z. Bai, M. D. Rosenberg, and Y. C. Leong, "Dynamic brain connectivity predicts emotional arousal during naturalistic movie-watching," *PLOS Comput. Biol.*, vol. 21, no. 4, p. e1012994, Apr. 2025, doi: 10.1371/journal.pcbi.1012994.
[23] J. Chen, Y. C. Leong, C. J. Honey, C. H. Yong, K. A. Norman, and U. Hasson, "Shared memories reveal shared structure in neural activity across individuals," *Nat. Neurosci.*, vol. 20, no. 1, pp. 115–125, Jan. 2017, doi: 10.1038/nn.4450.
[24] R. D. I. van Genugten and D. L. Schacter, "Automated scoring of the autobiographical interview with natural language processing," *Behav. Res. Methods*, vol. 56, no. 3, pp. 2243–2259, Mar. 2024, doi: 10.3758/s13428-023-02145-x.
[25] M. M. Bradley, M. K. Greenwald, M. C. Petry, and P. J. Lang, "Remembering Pictures: Pleasure and Arousal in Memory," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 18, no. 2, pp. 379–390, Mar. 1992, doi: 10.1037/0278-7393.18.2.379.
[26] J. A. Easterbrook, "The effect of emotion on cue utilization and the organization of behavior," *Psychol. Rev.*, vol. 66, no. 3, pp. 183–201, May 1959, doi: 10.1037/h0047707.
[27] D. J. Palombo, A. A. Te, K. J. Checknita, and C. R. Madan, "Exploring the Facets of Emotional Episodic Memory: Remembering 'What,' 'When,' and 'Which,'" *Psychol. Sci.*, vol. 32, no. 7, pp. 1104–1114, July 2021, doi: 10.1177/0956797621991548.
[28] E. Tulving, M. E. Le Voi, D. A. Routh, and E. Loftus, "Ecphoric Processes in Episodic Memory [and Discussion]," *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.*, vol. 302, no. 1110, pp. 361–371, Aug. 1983, doi: 10.1098/rstb.1983.0060.
[29] J. L. McGaugh, L. Cahill, and B. Roozendaal, "Involvement of the amygdala in memory storage: Interaction with other brain systems," *Proc. Natl. Acad. Sci.*, vol. 93, no. 24, pp. 13508–13514, Nov. 1996, doi: 10.1073/pnas.93.24.13508.
[30] M. T. R. van Kesteren, D. J. Ruiter, G. Fernández, and R. N. Henson, "How schema and novelty augment memory formation," *Trends Neurosci.*, vol. 35, no. 4, pp. 211–219, Apr. 2012, doi: 10.1016/j.tins.2012.02.001.
[31] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 1877–1901, Dec. 2020.
[32] M. Kosinski, "Evaluating large language models in theory of mind tasks," *Proc. Natl. Acad. Sci.*, vol. 121, no. 45, p. e2405460121, Nov. 2024, doi: 10.1073/pnas.2405460121.
[33] S. Michelmann, M. Kumar, K. A. Norman, and M. Toneva, "Large language models can segment narrative events similarly to humans," *Behav. Res. Methods*, vol. 57, no. 1, p. 39, Jan. 2025, doi: 10.3758/s13428-024-02569-z.
[34] J. Chen, "*Sherlock Movie Watching Dataset."* Oct. 26, 2016. doi: 10.34770/9ndy-8c50.
[35] J. Chen, Y. C. Leong, C. J. Honey, C. H. Yong, K. A. Norman, and U. Hasson, "Shared memories reveal shared structure in neural activity across individuals," *Nat. Neurosci.*, vol. 20, no. 1, pp. 115–125, Jan. 2017, doi: 10.1038/nn.4450.
[36] H. Lee and J. Chen, "Predicting memory from the network structure of naturalistic events," *Nat. Commun.*, vol. 13, no. 1, p. 4235, July 2022, doi: 10.1038/s41467-022-31965-2.
[37] M. Sakaki, K. Fryer, and M. Mather, "Emotion Strengthens High-Priority Memory Traces but Weakens Low-Priority Memory Traces," *Psychol. Sci.*, vol. 25, no. 2, pp. 387–395, Feb. 2014, doi: 10.1177/0956797613504784.
[38] L. J. Levine and R. S. Edelstein, "Emotion and memory narrowing: A review and goal-relevance approach," *Cogn. Emot.*, vol. 23, no. 5, pp. 833–875, Aug. 2009, doi: 10.1080/02699930902738863.
[39] S. F. C. Bartlett, *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press, 1932.
[40] W. F. Brewer and J. C. Treyens, "Role of schemata in memory for places," *Cognit. Psychol.*, vol. 13, no. 2, pp. 207–230, Apr. 1981, doi: 10.1016/0010-0285(81)90008-6.
[41] J. L. McGaugh, "The Amygdala Modulates the Consolidation of Memories of Emotionally Arousing Experiences," *Annu. Rev. Neurosci.*, vol. 27, no. 1, pp. 1–28, July 2004, doi: 10.1146/annurev.neuro.27.070203.144157.
[42] F. Dolcos, K. S. LaBar, and R. Cabeza, "Interaction between the Amygdala and the Medial Temporal Lobe Memory System Predicts Better Memory for Emotional Events," *Neuron*, vol. 42, no. 5, pp. 855–863, June 2004, doi: 10.1016/S0896-6273(04)00289-2.
[43] F. Dolcos, K. S. LaBar, and R. Cabeza, "Remembering one year later: Role of the amygdala and the medial temporal lobe memory system in retrieving emotional memories," *Proc. Natl. Acad. Sci.*, vol. 102, no. 7, pp. 2626–2631, Feb. 2005, doi: 10.1073/pnas.0409848102.
[44] A. Ehlers and D. M. Clark, "A cognitive model of posttraumatic stress disorder," *Behav. Res. Ther.*, vol. 38, no. 4, pp. 319–345, Apr. 2000, doi: 10.1016/S0005-7967(99)00123-0.
[45] C. R. Brewin, "Episodic memory, perceptual memory, and their interaction: Foundations for a theory of posttraumatic stress disorder," *Psychol. Bull.*, vol. 140, no. 1, pp. 69–97, Jan. 2014, doi: 10.1037/a0033722.
[46] O. Raccah, P. Chen, T. M. Gureckis, D. Poeppel, and V. A. Vo, "The 'Naturalistic Free Recall' dataset: four stories, hundreds of participants, and high-fidelity transcriptions," *Sci. Data*, vol. 11, no. 1, p. 1317, Dec. 2024, doi: 10.1038/s41597-024-04082-6.

Appendix

A. Large Language Model Prompt

The complete set of prompts and outputs for all events is provided in the GitHub repository:

https://github.com/xpan4869/llm-memory-scoring

While the exact wording varied, the minimal reproducible prompt structure for the task is illustrated below.

> *Task: Extract central details from movie scene annotations.*
>
> *Definition: Central details are causally essential elements of a narrative that sustain the storyline. They include information that drives the plot forward, explains character motivations, or marks turning points in the story. Without these details, the coherence or progression of the narrative would be disrupted.*
>
> *Inputs: {summary}, {annotation}*
>
> *Instructions: Identify only causally essential details; exclude descriptive context; express each in ≤10 words; ensure non-redundancy.*
>
> *Output: Table with ID + Idea Units.*

Peripheral detail prompts followed the same structure, with definitions adapted to emphasize descriptive but non-essential features. Scoring prompts instructed the model to assign 0/1/2 fidelity ratings to each detail.

B. Prior for Bayesian Multilevel Models

For each subject i, event j, and dataset k, the likelihood for an observed outcome $y_{i,j,k}$ is assumed to be drawn from a normal distribution:

$$y_{i,j,k} \sim Normal(\mu_{i,j,k}, \sigma)$$

where $\mu_{i,j,k}$ denotes the expected outcome for subject i, event j, and dataset k, and σ is the residual standard deviation of the outcome.

$\mu_{i,j,k}$ is modeled as a function of fixed intercept and fixed slope for the predictors with subject- and dataset-specific random intercepts:

$$\mu_{i,j,k} = \beta_0 + \beta_1 x_{i,j,k} + \alpha_{subj[i]} + \alpha_{dataset[k]}$$

with the following priors:

$$\beta_o \sim Normal(0, 1)$$

$$\beta_1 \sim Normal(0, 1)$$

$$\alpha_{subj[i]} \sim \text{Normal}(0, \sigma_{subj})$$

$$\alpha_{dataset[k]} \sim \text{Normal}(0, \sigma_{dataset})$$

$$\sigma_{subj} \sim Exponential(1)$$

$$\sigma_{dataset} \sim Exponential(1)$$

$$\sigma \sim Exponential(1)$$

These priors are weakly informative, regularizing estimates toward zero while allowing sufficient flexibility for variation across participants and datasets.