

AI-Enabled Human Memory Manipulation: Misleading AI-Generated Summaries Distort Human Memory

Mattea Sim¹, Yael Eiger², Tadayoshi Kohno¹

¹Georgetown University

²University of Washington

mattea.sim@georgetown.edu, yeiger@cs.washington.edu, yoshi.kohno@georgetown.edu

Abstract

AI-generated summaries and reports are increasingly used in high-stakes settings, like policing, despite considerable evidence that AI often generates misleading or inaccurate information. This research asked: do errors in AI-generated summaries distort human memory? To answer this question, we adopted two methodological approaches. First, we conducted an analysis of AI summary output, prompting popular large language models to generate summaries of videos. This analysis quantified how often AI summaries contain errors and the categories of these errors (e.g., omission of critical details; hallucinations), providing insight into the kinds of misleading information that may distort human memory. Second, we conducted a human-subjects experiment to empirically test the impact of misleading information in AI-generated summaries on human memory. Participants were first exposed to an event via watching a video of a car-pedestrian accident, and later read an AI-generated summary describing the video that either contained misleading or accurate information. Participants' memory for the original event was assessed in a memory recognition test. This work contributes several novel findings. In the AI analysis, we found a high frequency of mistakes in AI summaries, and particularly frequent omissions of critical details. For instance, the majority of summaries omitted the most central event of the video, a critical error which is likely to be impactful in high-stakes settings. We also observed a strong effect of AI misinformation on human memory. People who read a misleading AI summary were significantly less likely to accurately recall the original event, compared to people who read an accurate AI summary. These findings have implications for how AI should be used in critical settings. Though "humans-in-the-loop" are often expected to correct for AI's mistakes, our work suggests human memory can instead be distorted by these mistakes. AI has the potential to generate misinformation, even absent any adversarial intent, which can meaningfully impact human memory.

Data and Supplementary Materials —

<https://osf.io/wmjtn>

Introduction

Generative AI is being rapidly adopted and integrated for use in high-stakes settings, like policing (Potts et al. 2024),

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

hiring (Schellmann 2024), and healthcare (Doximity 2026). For instance, police departments can now automatically generate police reports from body-worn camera footage or audio (Kulkarni et al. 2025). However, these AI-generated reports or summaries can include mistakes. In one such example, an AI-generated police report mistakenly claimed a police officer shape-shifted into a frog after picking up on audio from "The Princess and the Frog" playing in the background (Tangermann 2026). To account for AI's fallibility in high-stakes settings, guidelines typically suggest human oversight, in which a person should check AI output to ensure the accuracy of the information (Potts et al. 2024).

Unfortunately, human memory is also fallible, and the assumption that people can notice and correct for errors in AI output may not always be true. In our work, we ask: how might misleading information in an AI-generated summary impact *human memory*? In fact, human memories can be quite susceptible to mistakes or misleading information. A seminal theory in the field of human memory, the *misinformation effect*, demonstrates that people's memories for events can be distorted by exposure to misleading information or misinformation after an event (Loftus, Miller, and Burns 1978; Loftus 2005).¹ Because AI can relay information (and misinformation) confidently and persuasively (Danry et al. 2025; Kabir et al. 2024; Kidd and Birhane 2023), and because these outputs are used in a variety of impactful settings (Potts et al. 2024; Doximity 2026), there is a real risk of AI creating misinformation at an unprecedented scale — and perhaps impacting human memory, as well.

We adopted two distinct approaches to explore the impact of AI summaries on human memory. First, we conducted an **analysis of AI summary output to quantify the frequency and the category of errors that appear in AI-generated summaries**. Here, we prompted popular large language models (LLMs) to summarize videos before quantifying *how often* errors occur and *how* errors occur — in other words, what categories or classes of errors arise in the output? Rather than conducting a systematic analysis of a

¹Throughout this paper, we use the term "misinformation" consistent with its use in the human memory literature, which refers to misinformation via both its definition as misleading or inaccurate information, *and* as the psychological effect misleading information has on human memory (the "misinformation effect").

wide range of videos, we conducted a more limited analysis of two brief, straightforward videos that are classically used in the human memory literature (Loftus, Miller, and Burns 1978; Wiechert et al. 2026) to more deeply understand the AI errors that might impact human memory. This analysis posed the following hypothesis:

- **H1:** AI-generated summaries will include a meaningful frequency of errors across a range of categories that could be impactful for human memory.

Second, we conducted a **human-subjects experiment to empirically test the impact of errors in AI-generated summaries on human memory**. Here, people witnessed an event (by watching the same classic memory videos used in the above analysis) before reading an AI-generated summary of the event that either contained accurate or misleading information about the event. People then completed a memory test assessing their recollection for the original event. Our central hypothesis was:

- **H2:** Misleading information in an AI-generated summary will distort people's memories for an event: i.e., an AI-enabled misinformation effect.

We also tested whether people's susceptibility to misinformation would depend on whether they believed the summary was AI-generated versus human-written. To do so, human participants were first told that the summary was written by *either* an AI model *or* a human (though in actuality, the summary was identical across conditions and was generated by ChatGPT). People also responded to questions assessing their trust in AI. This test posed three alternative hypotheses:

- **H3a:** People will be similarly susceptible to misinformation believed to be generated by AI and humans.
- **H3b:** People will be *more* or *less* susceptible to misinformation believed to be generated by AI than by a human.
- **H3c:** People's susceptibility to misinformation will depend on how much they trust AI.

This work makes several contributions. To foreshadow our findings, we observed that AI-generated summaries included a high frequency of errors across a range of categories. For instance, we found frequent omissions of critical details in each AI summary in our analysis, a type of error that can likely degrade human memory. Centrally, we also found that misleading information in an AI-generated summary distorted people's memory. People were susceptible to misleading information in AI summaries, regardless of whether they believed the summary was AI-generated or human-written, and regardless of how much they trusted AI or used AI. These findings have real implications for human-AI collaboration in high-stakes settings. Whereas humans are often positioned to account for and mitigate AI's mistakes, our work demonstrates that these mistakes in AI output could instead manipulate human memory, even for those who might be more skeptical of AI's output. By applying psychological theories of human memory to the rapidly evolving landscape of generative AI, we demonstrate both

human memory and AI are fallible, together painting a concerning picture for human-AI collaboration in critical settings where AI is increasingly adopted.

Background and Related Work

Misinformation and False Memories

For over 50 years, human memory researchers have explored the ways in which our memories are susceptible to manipulation (Loftus 1975, 2005). This research surfaced a phenomenon now known as *the misinformation effect*: people's recollection of events are often altered after exposure to inaccurate or misleading information (Loftus, Miller, and Burns 1978). In this classic study (and later replications) (Wiechert et al. 2026), people witnessed a car-pedestrian accident, in which a car comes to an intersection with either a stop sign or a yield sign. After, people were exposed to information about the traffic sign that was either accurate, misleading, or neutral (no mention). Finally, in a forced-choice memory recognition task, people had to recall whether they originally saw a stop sign or a yield sign. Those who were exposed to misleading information were considerably more likely to misremember the critical traffic sign detail.

The impact of misleading information can be quite powerful. Indeed, even a seemingly minor semantic detail, such as the strength of a verb used to describe a car crash (e.g., "smashed" versus "hit"), can alter later memory for related details like the speed of the cars (Loftus and Palmer 1974). More dramatic examples of misinformation include what are sometimes called "rich false memories," in which 25%–35% of adult participants were susceptible to a detailed false memory of being lost in the mall as a child (Loftus and Pickrell 1995; Murphy et al. 2023).

Once a false memory is implanted, it can also be quite robust and resilient. The effect of misleading information on human memory often persists even when people are directly warned that they were exposed to misinformation (Eakin, Schreiber, and Sergeant-Marshall 2003), and when people are told that everything they read after an event was inaccurate (Lindsay 1990). In other words, many people's memories are still swayed by misleading information, even when they are told to ignore the information (cf. Highhouse and Bottrill 1995). Once a false memory is implanted, it can be difficult to correct the memory later.

Though the misinformation effect is widely studied, technological advances facilitate novel ways of spreading misinformation and potentially impacting human cognition. In our work, we investigated how AI might facilitate a new iteration of the misinformation effect.

AI Misinformation and Its Impacts on People

It is now widely documented that LLMs (e.g., ChatGPT, Gemini) have a tendency to "hallucinate" (Rawte et al. 2023; Xu, Jain, and Kankanhalli 2024), perpetuate bias (Gueorguieva and Caliskan 2026; Nicolas and Caliskan 2025; Raj et al. 2026; Wilson and Caliskan 2024), omit relevant information (Oukelmoun et al. 2025), and make otherwise

inaccurate claims (Kabir et al. 2024; Zhou et al. 2023). Consequently, AI has the potential to spread impactful misinformation, even absent any adversarial intent.

A full review of the literature on LLM failure rates and categories is beyond the scope of our paper, but this work supports the idea that AI output frequently contains errors that could impact human memory. Researchers have surfaced, quantified, and taxonomized categories of errors in AI-generated summaries (Rawte et al. 2023; Xu et al. 2024). In research on dialogue summaries, for example, the output inaccurately ordered chronological events, missed important details, and made baseless inferences (Ramprasad, Ferracane, and Lipton 2024). Errors are also studied in higher-stakes contexts where AI-generated summaries are increasingly used. For instance, AI-generated medical summaries sometimes include hallucinations and omissions (Oukelmoun et al. 2025; Asgari et al. 2025), errors which could be highly impactful in a clinical context. Such mistakes persist in human-AI interactions beyond summarization. For instance, seeking information from ChatGPT often yields inaccurate responses (Kabir et al. 2024). Further, AI chatbots or companions are prone to highly sycophantic behaviors like flattering and agreeing with users, which can degrade the output's accuracy and reliability (Cheng et al. 2026; Fanous et al. 2025; Lee et al. 2026). Though we do not conduct a large-scale AI analysis to systematically study model behavior, questions which are thoroughly studied by these researchers and many others, our research is the first to begin to categorize AI summary misinformation (i.e., errors) in ways relevant to human memory effects.

In the face of considerable evidence for the fallibility of AI, increasing attention has turned toward understanding how AI interaction or exposure may be impactful for human cognition. For instance, exposure to biased simulated output from LLMs and text-to-image models can impact humans' own decision-making (Wilson et al. 2025) and implicit bias (Sim et al. 2025). Interactions with AI can even influence personal beliefs and attitudes more powerfully than typical interventions (Lin et al. 2025; Hackenburg et al. 2025). Indeed, persuasive messages generated by AI are often perceived as stronger than human-generated persuasive messages (Karinshak et al. 2023).

Of increasing interest is how AI-generated *misinformation* in particular might impact users (Kidd and Birhane 2023). People are often susceptible to this misinformation: for instance, when verifying answers to programming questions provided by ChatGPT, programmers overlooked nearly 40% of errors in ChatGPT's responses (Kabir et al. 2024). Chatbots can even influence people's political preferences by presenting relevant evidence and facts, including inaccurate and hallucinated claims (Lin et al. 2025; Hackenburg et al. 2025). Purposefully deceptive misinformation generated by AI can also be highly impactful and persuasive (Danry et al. 2025; Goldstein et al. 2024; Zhou et al. 2023). Relevant to the present work, AI misinformation can also impact human memory. Memories for news articles or photos were influenced by interacting with a misleading chatbot (Pataranutaporn et al. 2025a) or exposure to AI-edited images (Pataranutaporn et al. 2025b). Building on this

work, we study how misleading AI-generated summaries, increasingly used in a variety of contexts, may impact human memory.

A related and novel question investigated in our work is whether people's susceptibility to misinformation is impacted by the belief that the information was human-written versus AI-generated. Information source can impact human perception: people tend to prefer AI-generated information *unless* it is explicitly labeled as AI-generated (Karinshak et al. 2023). Further, people are aware that AI-generated information contains errors, and thus are more skeptical of its output (Kabir et al. 2024). Yet users also increasingly rely on AI for work, information-gathering, and other high-stakes tasks for which accuracy is key (Angrisani et al. 2026; Milella and Cabitza 2026), suggesting at least some users see AI as a highly reliable source. Further, sources that appear confident and objective, like AI, are more readily trusted (Kidd and Birhane 2023). Therefore, we explored whether human memory's susceptibility to misleading information depends on the supposed source: AI or human.

Methodology

Our methodology consisted of two parts. First, we conducted an analysis of popular AI models with video-to-text summarization capabilities to quantify the number of AI output errors and the category of these errors. Second, we conducted a human-subjects experiment in which we tested the degree to which an error in an AI-generated summary would distort people's memories — that is, an AI-enabled misinformation effect.

AI Summary Analysis Procedure

We conducted an analysis to quantify and categorize the errors produced in AI-generated video summaries with two of the most popular large language models with video-to-text summarization capabilities: ChatGPT-5.5 (OpenAI 2026) and Gemini 2.5 Flash-Lite (Gemini Team 2025). We focused the analysis on the two 25-second videos that were also presented to participants in the human-subjects study (see Human Memory Study Procedure below), both of which were used in prior memory studies and depict an animated car-pedestrian accident (Wiechert et al. 2026). Rather than conducting a large-scale analysis of many videos, we conducted a smaller analysis of these two videos, prompting each model five times per video. This analysis allowed us to use the same videos across both the AI analysis and the human study, so that our work is consistent and deeply grounded in the human memory literature (Loftus, Miller, and Burns 1978; Wiechert et al. 2026). This also allowed us to investigate the *consistency* of models across time — that is, do models consistently make the same errors, or do models make different errors randomly across repeated tests? The videos were also a conservative test of errors: videos were short, simplistic, and presented minimal details or events, which may actually *reduce* the likelihood of errors and make any resulting errors more striking. This process led to 20 total video summaries which were qualitatively coded for errors.

Videos. The two videos were created in a recent replication study (Wiechert et al. 2026), which modernized materials from a classic human memory study paradigm (Loftus, Miller, and Burns 1978). We selected these videos because of the significance of this study paradigm in the human memory literature, and because the videos were publicly available online. The videos showed an animated depiction of a car-pedestrian accident created in Grand Theft Auto V (Rockstar Games 2013). The two videos were similar in content with the exception of a traffic sign: the videos show a red car approaching an intersection with *either* a stop sign or a yield sign. Both videos show the red car turning right at an intersection as a pedestrian walks into the road. The car collides with the pedestrian, who falls to the ground. The pedestrian stands up, the driver of the car exits the vehicle, and the driver and the pedestrian meet.

Prompting the Models. We chose to use a single prompt across the AI analysis because the goal of this work was not to systematically study how different prompts change the output of LLMs, questions that other researchers are studying deeply (Cantini and De Mauro 2026; Marvin et al. 2024; Schulhoff et al. 2025). Instead, our goal was to select a high-quality prompt with which we could quantify the different kinds of errors arising in LLM output.

In order to select the prompt, we conducted a series of preliminary tests trying different prompts and models for video summarization. For instance, we experimented with example prompts designed to summarize body cam footage (Adams 2024). We also prompted ChatGPT to generate a prompt idea based on our basic experimental design, which we tested. These prompts all included explicit instructions to maintain neutrality, be factual, and avoid making inferences. When testing prompts that omitted neutrality instructions, we observed what appeared to be qualitatively different summaries as a result, often lower quality with more erroneous assumptions. In order to adopt a conservative approach with reasonable, higher-quality instructions, our prompt included direction to generate neutral and factual summaries.

Further, we included a word minimum to generate summaries of fairly consistent lengths (at least 300 words), based on some existing guidelines for standard-length AI-generated reports (Axon 2026). We also conservatively added instructions to ensure everything important is included in the summary, due to a high frequency of notable omissions observed in initial testing. These prompt choices were intended to give the model its “best chance” at generating accurate summaries.

The final prompt reads:

I am seeking your expertise in generating a narrative from a video. Generate a factual, neutral text summary that is at least 300 words and provides a coherent narrative of everything that happens in this video. Do not introduce new facts, speculation, or interpretations. Make sure everything important is included in the summary. Write in third person, past tense. Produce only the summary text.

All final queries were done in Spring 2026. To ensure re-

peated prompt responses were not affected by prior memory of the video, each query was run using a fresh upload of the video and prompt each time, on fresh instantiations of the model, five separate times. This created a corpus of five responses from the yield sign video with Gemini, five responses from the stop sign video with Gemini, five responses from the yield sign video with ChatGPT, five responses from the stop sign video with ChatGPT — or 20 total summaries; 10 for each video; and 10 for each model. The responses from the model were saved to a CSV.

Coding Analysis. In order to code the AI summaries, three researchers first independently coded videos for artifacts before meeting to refine and standardize a list of what we considered to be the “central details” of the event. The list was generally intended to capture the “who,” “what,” “when,” “where,” and “why” of the event. The final list included 19 central details.

Prior to coding the summaries, we agreed upon four different classes or categories of errors to analyze based on our initial testing: 1) *omissions*, or failure to mention one or more of the 19 central details; 2) *critical inaccuracies*, or errors describing the 19 central details; 3) *non-critical inaccuracies*, or errors unrelated to the central details; and 4) *additions*, or hallucination-like addition of details not present in the event. Each error could only be coded in a single category.

Two researchers independently coded each resulting summary to quantify and categorize the errors. To do so, both researchers coded each independent occurrence of a central detail omitted, a central detail described inaccurately, a non-central detail described inaccurately, and hallucinated additions, resulting in “counts” within each class of error for each summary. Whereas the first two categories were coded based on the pre-determined central details list, the latter two categories were coded without pre-determining a list of non-central details or potential hallucinations; in other words, researchers simply coded any other inaccuracies or additions in the summaries that were inconsistent with the videos. The researchers met to resolve disagreements, reviewing the videos to reach a final consensus.

Human Memory Study Procedure

Participants completed a two-part study online via Prolific. The experimental design was a conceptual replication and extension of a classic memory study paradigm in which participants see a car-pedestrian accident and later read a summary of the accident that manipulates whether the car came to a stop sign or a yield sign, a critical detail which participants are then asked about in a memory recognition task (Loftus, Miller, and Burns 1978; Wiechert et al. 2026). We retain these core elements of the paradigm and extend it to investigate whether AI-generated summaries produce a misinformation effect amongst human participants, *and* whether participants’ memories are further influenced by believing that the summary is either AI-generated or human-generated. See Appendix for the full materials and measures.

Part 1: Encoding the Original Event. In the first part of the study, participants were told they would be acting as

Gender	%	Age	%	Race/Ethnicity	%
Man	51.8	18-24	5.2	White/European American	77.1
Woman	46.6	25-34	23.8	Asian/Asian American	10.4
Non-Binary	1.2	35-44	32.3	Black/African American	9.5
		45-54	23.5	Hispanic/Latino	6.1
		55-64	7.9	Multiple Identities Selected	5.2
		65+	7.3	Native Hawaiian/Pacific Islander	0.3

Table 1: Participant self-reported demographic frequencies.

an eye-witness to an animated car-pedestrian accident and were asked to pay close attention to the video. Participants watched one of two 25-second animated videos created in Grand Theft Auto V by researchers of the original study; videos are available online via their paper (Wiechert et al. 2026). The videos showed a red car approaching an intersection with *either* a stop sign or a yield sign. The traffic sign was counterbalanced across videos (i.e., people were randomly assigned to see one of the two signs) to rule out potential stimulus effects, such as if one traffic sign was particularly memorable and created unique effects. Otherwise, both videos showed the same events with differences in only peripheral visual details (see description in “Videos” section of AI Summary Analysis Procedure). After participants watched the video, they answered an attention check question which asked the color of the main car in the video to ensure they paid attention. Only participants who correctly answered “red” were invited to take the second part of the study.

Part 2a: Post-Event Summary. We implemented a 24–48 hour gap between the two parts of the study, consistent with findings that the impact of misleading information on memory is stronger after a longer time delay (Loftus, Miller, and Burns 1978; Wiechert et al. 2026). Twenty-four hours after watching the video, participants who passed the attention check were invited to take part in the rest of the study via Prolific, which was live online for the subsequent 24 hours or until all participants completed the study.

In the second portion of the experiment, participants read a 21-sentence AI-generated text summary describing the details of the video.

The AI-Generated Summary. In our experimentation with video summarization, ChatGPT often produced summaries with errors, such as the omission of critical details, and these errors were often irregular or inconsistent across responses. Thus, one researcher (on a separate account from the AI summary analysis prompting) repeatedly re-prompted ChatGPT, ultimately selecting one of the higher-quality summaries returned to show participants (with minor modifications to ensure accuracy, described below; see Appendix for the full original and modified summary).² While

²After noticing prompts were omitting details of the car-pedestrian accident, one researcher used a new ChatGPT account (separate from the account used in the AI analysis) to re-prompt the model until the accident was included in the resulting summary. Subsequently, in some cases, ChatGPT would return accurate details, whereas in other cases, it would not. We consider it beyond

this re-prompting process may not accurately represent all uses for ChatGPT-based summarization (i.e., we already knew the central details of the video and were striving to ensure the AI summary included those details), our process ensured that the summary used in this part of the study was a valid ChatGPT output.

The final summary was generated using the same prompt as in the AI Summary Analysis Procedure, and summarized the yield-sign video. There were slight peripheral differences between the yield video and the stop video, such as different advertisement boards and surrounding buildings. However, we chose to use one consistent summary for all participants in order to avoid confounding differences across conditions. Indeed, generating two separate summaries for different videos in ChatGPT often yielded meaningfully different text, with variability in descriptions and errors.

Thus, we made only minor edits, such as removing the specific details that were inconsistent across the yield and stop videos so that the summary would be accurate regardless of which video participants watched (e.g., removing descriptions of specific ad signs), and adding minor text that had been previously generated by ChatGPT. For instance, in prior experimentation, AI summaries of the stop sign video described it as a “red stop sign,” and thus, this AI-generated text was injected back into the summary for participants who were assigned to read the stop sign summary.

Experimental Conditions. Prior to reading the summary, participants were first randomly assigned to one of two “Summary Conditions,” in which they read that the summary was written by either an AI software designed to professionally transcribe videos (*AI Condition*), or a person who works as a professional video transcriber (*Human Condition*). In actuality, the summary was identical for all participants (with the exception of one critical traffic sign detail) and was generated by ChatGPT.

In addition, participants were randomly assigned to one of two “Information Conditions,” in which the summary included a description of the traffic sign that was either consistent with the sign from the video (*Consistent Condition*) or inconsistent with the sign from the video (*Misleading Condition*) — in other words, the summary described either a stop sign or a yield sign, either of which could be consistent or misleading depending on the original video. Though we controlled the presence of the traffic sign description to experimentally manipulate misinformation, we did so using

the scope of this study to diagnose exactly when and how ChatGPT can evolve to produce better summaries.

actual language output by ChatGPT, so that all text included in the summary was originally AI-generated.

Part 2b: Remembering the Event. Immediately after reading the summary, participants completed filler questions assessing the quality of the summary. Participants were then instructed to think back to the original video event (*rather than* the AI- or human-generated summary) to answer a series of questions, emphasizing their answers should be based on their own recollection of the video.

Participants responded to a 12-item memory recognition task with two-choice response options based on the details of the video. Ten of these questions were taken directly from prior work (Wiechert et al. 2026). Of central interest was the critical question, asking at which traffic sign the car came to a hold at the intersection. Participants could select either “stop” or “yield.”

Participants responded to a 10-item scale assessing trust in AI (Wang et al. 2026), and reported their frequency of AI use and other basic demographics before being debriefed.

Participants. All participants were recruited from Prolific and compensated at the minimum wage rate of the first author’s city, which is higher than the federal minimum wage. We relied on an a priori power analysis conducted in G*Power (Faul et al. 2007) from prior work (Wiechert et al. 2026), which determined that at least 63 participants per condition would provide 80% power to detect a medium effect size in a Chi Square test ($w = 0.25$, $df = 1$). We over-sampled to account for participant drop-off and attention check failures, collecting data from 360 adult U.S.-based participants in part 1. Using Prolific’s available screeners, participants were required to be fluent in English and to report having no issues seeing colors. Five of these 360 participants were not invited to participate in the full study due to a failed attention check (but were still paid for their participation in part 1).

In total, 331 participants completed both parts of the study. Three of these participants failed an attention check in part 2, leaving 328 participants included in analyses (AI Consistent Condition $n = 82$; AI Misleading Condition $n = 80$; Human Consistent Condition $n = 83$; Human Misleading Condition $n = 83$). See demographics in Table 1.

Results

All data and supplementary appendix materials are available online (<https://osf.io/wmjtn>) for reproducibility and open science.

AI Summary Analysis Results

We first quantified and categorized errors in AI summary output (H1). We connected each category of error to findings in the psychology literature that suggest the error could likely be impactful for human memory.

AI Summary Errors Are Frequent and Irregular. We first tested H1 to quantify the frequency and category of errors in AI-generated summaries of our test videos.³

³See Appendix for a table including the counts of each error category occurring across all twenty individual summaries (i.e., across

All twenty summaries included mistakes. Table 2 shows the percentage of summaries that included each category of error. We found a *minimum* of seven errors in each summary, up to a maximum of twenty-one errors in a given summary — notably high in summaries that averaged just 305 words. Both ChatGPT and Gemini made mistakes, though the mistakes themselves descriptively varied. For instance, whereas ChatGPT output included more errors of “addition,” Gemini output included more errors when describing central details of the event. In short, both AI models were prone to generating mistakes when summarizing videos.

The mistakes themselves were often inconsistent across summaries. For instance, a car would be omitted in one summary, but included in the next; hallucinated vehicles or pedestrians would appear spuriously; the main car would be described as turning right in one summary, and left in the next. Although certain errors were particularly common (e.g., omissions described in the next section), overall, errors in AI summaries were not consistent or uniform.

AI Summaries Omit Critical Details. What *kinds* of mistakes arose most often in AI summaries? The most common category of error was omission, or failure to mention the primary events of the video. Unlike the other categories of errors, omission errors were present in every summary.

Table 3 shows the percentage of *central details* across all summaries that were omitted, inaccurate, or accurate (whereas Table 2 shows the percentage of summaries with each category of error, Table 3 shows the percentage of central details that were omitted altogether or misconstrued in some way). Across all prompts and models, the summaries omitted 51.6% of central details on average.

Most notably, all but one summary (95% of summaries) omitted the single most central detail of the event: the car colliding with the pedestrian. 90% of summaries also omitted related key details of the event, in which the pedestrian is lying on the ground and stands up after the collision. Some summaries even omitted the pedestrian altogether, such as one summary that goes so far as to say “no other characters or vehicles interacted directly with the scene.”

Result: AI summaries omitted critical details.

Connection to Human Memory: Human memory is often degraded when accurate details are omitted from (human-written) text (Loftus, Miller, and Burns 1978).

AI Summaries Misconstrue Critical and Peripheral Details. AI summaries also included a notable number of critical inaccuracies, or errors when describing the 19 central details of the videos. 65% of summaries inaccurately described at least one central detail. Summaries contained an average of 1.15 errors related to central details, or 6.1% of central details described inaccurately across summaries. For instance, summaries often inaccurately described interactions between different elements of the scene (e.g., a car passing around another car), patterns of movement (e.g., a

the two models and repeated prompts for the stop and yield videos).

	Omissions	Critical Inaccuracies	Non-Critical Inaccuracies	Additions
ChatGPT	100%	30%	80%	70%
Gemini	100%	100%	100%	50%
Total	100%	65%	90%	60%

Table 2: Percent of summaries that included at least one of each category of error.

	Omissions	Inaccuracies	Accuracies
ChatGPT	52.1%	3.2%	44.7%
Gemini	51.1%	9%	40%
Total	51.6%	6.1%	42.4%

Table 3: Percent of *central details* omitted, inaccurate, or accurate across all summaries in each model.

car turning right, a pedestrian walking into the road), and chronological sequences of events (e.g., a car stopping before a driver exits the car). In our analysis, Gemini appeared more prone to this category of error compared to ChatGPT.

AI summaries misconstrued more peripheral or non-central details of the videos, as well, averaging 2.05 non-critical inaccuracies per summary. For instance, summaries often inaccurately described visual elements of the scene like the text on signs or positions of the cars.

Result: AI summaries inaccurately described important and minor details.

Connection to Human Memory: Inaccurate descriptions of both central and peripheral details in (human-written) summaries can distort memory (O'Donnell and Chan 2025).

AI Summaries Hallucinate and Directly Contradict or Alter Central Details. Finally, AI summaries also included notable hallucinations or additions of details not included in the videos, with an average of 1.6 errors per summary. Over half of the summaries (60%) included at least one hallucination or addition. For instance, summaries hallucinated additional pedestrians and vehicles, erratic vehicle movements, and inferences about characters' intentions, all of which meaningfully alter or misconstrue important events in the video.

Some AI summaries included errors that directly contradicted important details. For instance, one summary stated that the pedestrian slightly repositions but does not enter the roadway, whereas the video shows the pedestrian walking into the road, a key sequence of events which causes the car-pedestrian accident.

Result: AI summaries added erroneous details.

Connection to Human Memory: Inaccurate additions to an event in (human-written) text can implant false memories (Loftus 1975).

Human Memory Study Results

We next *empirically* tested the claim that an error in an AI summary could distort human memory (H2), with data from the human-subjects experiment. We further investigated whether human susceptibility to misinformation depended on the supposed source of information or people's trust in AI (H3).

A Misleading AI Summary Distorts Human Memory.

We tested our central hypothesis (H2) that AI summaries containing misleading information would impair people's memory for that information. There was a significant relationship between the Information Condition and the likelihood of answering the critical traffic sign question correctly, $\chi^2 = 53.937$, $p < .001$. Consistent with our hypothesis, 83.6% of participants who read a consistent summary answered the critical question correctly, compared to only 44.8% of participants who were correct after reading a misleading summary.⁴ These results demonstrate a strong AI-enabled misinformation effect, whereby misleading AI summaries distort memory.

Memory Finding: Misleading information in an AI summary can distort people's memory for an event.

People are Similarly Susceptible to AI and Human Misinformation Regardless of Their Trust in AI.

We next tested whether susceptibility to misleading information would depend on whether people believed the summary was written by an AI model versus a human (H3). There was no significant relationship between the Summary Condition (AI- versus human-labeled) and the likelihood of answering the critical question correctly, within either the consistent condition, $\chi^2 = 1.181$, $p = .277$, $\phi = .085$, or the misleading condition, $\chi^2 = 0.136$, $p = .712$, $\phi = -.029$.

In other words, there was a strong misinformation effect *regardless* of how people believed the summary was written, indicated by significant relationships between the Information Condition and the likelihood of answering the critical question correctly, within both the AI-labeled summary condition, $\chi^2 = 20.499$, $p < .001$, $\phi = -.356$, and the

⁴The misinformation effect was strong in magnitude regardless of whether participants originally saw the stop sign video, $\chi^2 = 37.995$, $p < .001$, $\phi = -.481$, or the yield sign video, $\chi^2 = 25.002$, $p < .001$, $\phi = -.390$, though overall accuracy was higher for those who saw the stop sign video. Prior work found people do encode the traffic sign to memory without summaries, with 80.26% of people correctly recalling the traffic sign after watching the video (Wiechert et al. 2026).

human-labeled summary condition, $\chi^2 = 34.345$, $p < .001$, $\phi = -.455$ (see Figure 1).

Descriptively, amongst participants who read misleading summaries, just 46.3% in the AI condition and 43.4% in the human condition answered the critical question correctly. Human memory was strongly susceptible to misinformation, regardless of whether the summary was supposedly written by a human or an AI model, supporting Hypothesis 3a.

Memory Finding: Misleading AI summaries can consistently distort human memory, regardless of whether summaries are supposedly human-written or AI-written.

We also explored whether participants' level of trust in AI or frequency of using AI would moderate the extent to which they were susceptible to misinformation across the AI and human conditions. We first created a composite AI Trust variable averaging across 10 items ($\alpha = .959$; $M = 3.10$, $SD = 0.97$; higher numbers = more AI trust).

We conducted a binary logistic regression to investigate the interaction between Summary Condition (Human = 0, AI = 1), Information Condition (Consistent = 0, Misleading = 1), and AI Trust (mean-centered) on the likelihood of answering the critical traffic sign question correctly (0 = incorrect; 1 = correct). There were no significant interactions ($ps > .16$), including the central three-way interaction, $\beta = -0.131$, $SE = 0.577$, $p = .820$, $\exp(\beta) = .877$, nor was there a significant relationship between AI trust and the likelihood of answering the critical question correctly, $\beta = -0.344$, $SE = 0.381$, $p = .366$, $\exp(\beta) = .709$.

Running the same binary logistic model but with participants' reported frequency of AI chatbot use (mean-centered; $M = 3.06$, $SD = 0.93$) replacing AI trust, the three-way interaction was not significant, $\beta = -0.240$, $SE = 0.397$, $p = .547$, $\exp(\beta) = .787$, nor was there a significant relationship between frequency of AI use and the likelihood of answering the critical question correctly, $\beta = -0.068$, $SE = 0.188$, $p = .718$, $\exp(\beta) = .934$.

Memory Finding: The extent to which people trust AI or use AI does not change their susceptibility to misleading AI summaries.

Therefore, the susceptibility of human memory to misinformation did not depend on whether people believed the summary was written by an AI model or a human, nor did it depend on trust or familiarity with AI.

Non-Critical Questions. Finally, we explored several non-critical questions. We tested accuracy on non-critical memory questions to address the possible critique that participants had no memory of the original video event, and were instead relying entirely on the text summary. We tested accuracy on a question which was *not* mentioned in the summary: namely, the time at which the accident occurred (day versus night). 96.6% of participants correctly identified that the accident occurred in the daytime, providing strong evi-

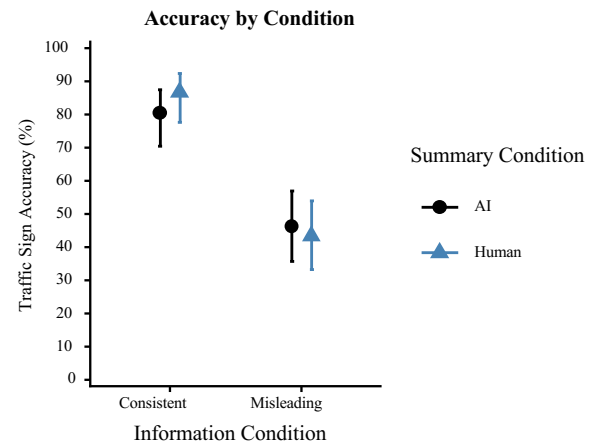


Figure 1: Percent of participants who answered the critical question correctly across conditions. Vertical error bars represent 95% confidence intervals.

dence that participants were able to recall the original event. On average, participants answered correctly to 91.4% out of 11 non-critical questions.

Although not central to our hypotheses, we also investigated whether participant ratings of summary clarity, readability, or enjoyment would differ across conditions. There were no significant effects of Information Condition ($ps > .410$, $n_p^2s < .003$), nor significant interactions between Summary Condition and Information Condition ($ps > .290$, $n_p^2s < .004$), on any of these variables. However, there was a significant effect of Summary Condition on perceived clarity of the text summary, $F(1, 326) = 7.800$, $p = .006$, $n_p^2 = .024$. Participants believed the human-labeled summary was more clearly written ($M = 9.01$, $SD = 1.07$) than the AI-labeled summary ($M = 8.62$, $SD = 1.47$), $t(326) = -2.792$, $p = .006$, $d = -0.31$, $95\%CI[-0.53, -0.09]$ (though in actuality, the summaries were identical). Participants also rated the human-labeled summary as descriptively (but not significantly) more readable ($M = 9.16$, $SD = 1.14$) than the AI-labeled summary ($M = 8.97$, $SD = 1.34$), $p = .159$, $d = -0.16$.

Discussion

The extent to which mistakes in AI summaries impact human cognition is critical to understand as AI is increasingly used to summarize important information in real-world contexts. In this work, we adopted two distinct methodological approaches to investigate errors in AI summaries and their impact on human memory. Centrally, we conducted a human-subjects experiment simulating human-AI interaction in which we empirically demonstrated that misleading information in an AI summary distorted people's memories for an event (H2). Given the widespread use of LLMs in both organizational contexts and amongst daily individual users (Angrisani et al. 2026; Doximity 2026; Milella and Cabitza 2026; Potts et al. 2024), our findings suggest the potential for AI to create misleading information at scale —

with real consequences for human memory.

We also found that people were susceptible to misinformation in AI summaries *regardless* of whether they believed the summary was written by AI or a human (H3). This finding extends our theoretical understanding of the misinformation effect, a phenomenon studied in the psychology literature for over 50 years (Loftus 1975, 2005). Misleading information, as classically written or administered by humans, has a robust impact on human memory. Our findings contribute to a recent wave of work on AI and memory (Pataranutaporn et al. 2025a,b), which extends the misinformation effect to generative AI technologies, and we provide novel evidence that the effect may be robust to the apparent source of misinformation. Indeed, people are often more skeptical of information believed to be generated by AI (Kabir et al. 2024; Karinshak et al. 2023), though many people also increasingly rely on AI for information (Angrisani et al. 2026; Milella and Cabitza 2026). Despite potentially polarizing opinions about AI, we found that the misinformation effect is strong in magnitude, regardless of whether the source is believed to be AI or human.

Even further, the impact of misleading AI summaries on human memory did not depend on people's trust or familiarity with AI. These results are particularly striking: even those who are critical of AI may be susceptible to an AI-enabled misinformation effect. In other words, an awareness that AI makes mistakes may not be enough to combat the effect of misleading information in an AI summary on human memory. Yet AI frequently generates mistakes, biases, and hallucinations (Rawte et al. 2023; Wilson and Caliskan 2024; Kabir et al. 2024), and in some contexts, people collaborating with AI are expected to notice and mitigate these errors (Tabassi 2023). Though we interpret a non-significant statistical finding with caution, our results suggest that solutions like increasing AI literacy alone may not be enough to combat these effects.

AI Summary Errors in Human Memory Context

We also conducted an analysis of AI summary output for videos that are significant to the human memory literature, illustrating a high frequency of errors that are likely to impact human memory (H1). A significant body of work is focused on detecting, categorizing, and minimizing errors in LLM output (Ramprasad, Ferracane, and Lipton 2024; Rawte et al. 2023; Xu et al. 2024; Xu, Jain, and Kankanhalli 2024). Our findings contribute to this literature and the literature on human-AI interaction by connecting common classes of errors in AI summaries to psychological memory effects, suggesting the errors could have a real impact on people. For instance, the most common category of AI summary error was *omission*, in which the summary failed to mention central details of the videos. This type of error has been empirically observed in higher-stakes AI-generated summaries such as those used in medical settings (Oukelmoun et al. 2025), and also has significant connections to the human memory literature. Some studies of the misinformation effect include a “no information” condition, in which people are exposed to summaries or post-event information that omits a critical detail altogether. Though typically con-

sidered a control condition, people are less able to accurately recall critical details after exposure to *no information* as compared to accurate information (Loftus, Miller, and Burns 1978). Therefore, if AI-generated summaries omit important details, this error may have impactful consequences for human memory.

In fact, all categories of errors we observed in AI-generated summaries may be likely to impact human memory. In our human-subjects experiment, we implanted a false detail into an AI summary that is akin to a *critical inaccuracy*, or an inaccurate description of a central detail of the video. This is perhaps one of the most commonly studied types of misinformation in the human memory literature, along with inaccuracies related to more peripheral details (O'Donnell and Chan 2025), with robust effects on human memory (as also demonstrated in our study). Relevant to the error of hallucinations, which are frequently studied in analyses of LLM behavior (Xu, Jain, and Kankanhalli 2024), other studies demonstrated that “false presuppositions,” in which people suggest the existence of objects that were not actually present in an event, can implant false memories of those objects (Loftus 1975). Thus, hallucinations or erroneous additions in AI summaries may be able to manipulate human memory, as well.

Implications for AI Summaries in Practice

Broadly, our findings have implications for how AI should be used in real-world settings. As generative AI is increasingly aiding humans in high-impact decisions, most guidelines stress the importance of “humans-in-the-loop” to collaborate with AI and ensure the accuracy or appropriateness of final decisions (Tabassi 2023). In other words, people are expected to notice and mitigate potential biases or errors in AI output. However, our work and other recent work calls into question the effectiveness of this approach. For instance, when collaborating with AI to make hiring decisions, people tend to over-rely on AI recommendations, even when these recommendations are racially biased (Wilson et al. 2025).

Our work contributes to growing evidence that the human-in-the-loop strategy is not enough to account for AI's mistakes. That errors in AI-generated summaries distort human memory paints a particularly concerning picture when considering how AI may be used to summarize information in real-world contexts, like policing. Though outlandish mistakes like an officer turning into a frog may be more easily corrected (Tangermann 2026), our work demonstrates that AI summaries could include a range of more subtle (yet still impactful) errors, and these errors can alter human memory.

Notably, policing has been at the forefront of discussions around the misinformation effect for as long as the effect has been studied in psychology. The misinformation effect is often connected to eyewitness testimony, in which misleading questions or interrogations can manipulate eyewitness memory (Loftus 1975; Loftus and Klemfuss 2024). Now, the use of AI in policing and related settings raises new questions about how the misinformation effect might manifest in a legal context. Technological advances can change and exacerbate the influence of misinformation, problems worthy of continued research and deep consideration as AI is adopted

into these critical settings.

Finally, although a great deal of work in the area of misinformation studies *deliberate* disinformation campaigns (Goldstein et al. 2024; Zhou et al. 2023), in which adversaries may seek to spread misinformation to manipulate or cause harm, the present work highlights a potential harm that could arise *with or without* an adversary. AI generates misleading information without malicious intent, and this misinformation can distort human memory. Together, this presents a real risk of an AI-enabled misinformation effect, whereby AI can widely create misleading information and consequently impact human cognition.

Open Questions and Limitations

The current work had limitations that may motivate future work. First, we conducted a more limited analysis of AI model output, which constrains the conclusions we can make about LLM behavior. This analysis was based on two videos previously utilized in the memory literature depicting video game footage of a simple car accident, which limits the generalizability of this work to other more realistic videos. Further, we used a single standardized prompt and analyzed a smaller number of resulting summaries. The goal of our research was not to systematically study all errors in LLM behavior, but rather to assess and illustrate how errors manifest in videos grounded in the human memory literature. That we found errors with both ChatGPT and Gemini does suggest the possibility for errors could be systemic: in the summaries of two videos, both AI systems had significant issues. However, we encourage future research to study a higher volume of realistic videos and summaries to better understand these errors, while also pointing to a significant existing literature that more deeply studies current LLM behavior (Ramprasad, Ferracane, and Lipton 2024; Rawte et al. 2023; Xu et al. 2024; Xu, Jain, and Kankanhalli 2024).

Second, we tested whether people's susceptibility to misinformation in a video summary would differ depending on whether they believed the summary was written by AI versus a human. However, in actuality, all participants read an identical summary generated by ChatGPT. This design choice provided us with more experimental control, whereby any differences observed across conditions could be attributed to the manipulation, rather than to confounding differences in summary text. One possible weakness of this design is participants in the human condition could have suspected that the summary was actually AI-generated, reducing potential differences across the AI and human conditions. However, if true, we would not expect to see significant differences in the perceived clarity of the identical summary across human-labeled and AI-labeled conditions. Even further, participants were similarly susceptible to supposedly human or AI misinformation regardless of their trust in AI or how often they reported using AI, suggesting human memory may be similarly fallible regardless of the source.

Third, while video summarization can be a critical task in some contexts, some real-world use cases may summarize other forms of media, like audio transcripts (Kulkarni et al. 2025). It will be important for future work to investigate how different forms of media may impact the accuracy and reli-

ability of AI summaries. In fact, one hypothesis is that summaries created from audio transcripts alone could be *less* accurate, given the wealth of visual contextual information that would be missing from such transcripts. The methodological choice to summarize videos may have served as a more conservative test of the errors in AI summary output, but future work could study context-dependent methods of AI summarization.

Fourth, we conducted a highly controlled human-subjects experiment to expose people to an event and subsequent misinformation, following typical procedures in the psychological memory literature. However, complementary future work may benefit from exploring how the misinformation effect manifests in more naturalistic studies conducted within real high-stakes settings. Further, the conclusions of our work are limited by studying only one type of misinformation in simple animated car accident videos. Future iterations of these studies could include a range of frequently occurring errors (e.g., omissions of critical details) based on more realistic videos or events to further understand the impact of AI errors on human cognition.

Inspired by our findings, another question to explore in future work is whether or how to intervene during human-AI collaboration. If AI summaries are used in high-stakes settings like policing, are there strategies that can be implemented to help officers resist the influence of misleading information? For instance, noticing or detecting discrepancies between the original event and later misinformation can reduce the impact of misleading information on human memory (Putnam, Sungkhasettee, and Roediger III 2017). Warning people about potential discrepancies or misinformation prior to reading AI-generated summaries may help prompt people to notice errors (Greene, Flynn, and Loftus 1982; Loftus and Klemfuss 2024). However, warnings are not always effective (Loftus 2005), and human memory can still be susceptible to misinformation after such warnings (Eakin, Schreiber, and Sergeant-Marshall 2003). Given the robust effect of misleading information on human memory in our findings, and the critical decisions at stake, a deeper question to consider is whether AI should be used in these contexts at all.

Conclusion

Taken together, we found that errors in AI-generated summaries can distort human memory. The present work contributes to a growing body of research suggesting AI has real consequences for human cognition. Further, these findings have implications for how AI should be used in high-stakes settings, where humans who are meant to correct for AI's mistakes might instead be susceptible to false memories.

Ethical Statement

The study was approved by the human subjects review boards (IRBs) of all authors' institutions. Participants were informed that the full purpose of the research would be provided after completing the study; participants were then debriefed after completing the study. We took additional precautions. For example, while our final video selection was

driven by videos used in prior studies of human memory, in our initial (unrestricted) exploration of possible videos, we found (and then chose to exclude) videos with graphic content, such as real police body cam footage.

Researcher Positionality Statement

The authors' backgrounds influence the way we approach these research questions. We are a cross-disciplinary research team with backgrounds in AI ethics, computer security and privacy, and social psychology.

Acknowledgments

We are grateful to the reviewers for their feedback to improve the paper. This research was supported by the Robert L. McDevitt, K.S.G., K.C.H.S. and Catherine H. McDevitt L.C.H.S. Chair in Computer Science at Georgetown University and by the U.S. National Science Foundation under Award #2205171.

References

- Adams, I. T. 2024. Large Language Models and Artificial Intelligence for Police Report Writing. *CrimRxiv*. <https://www.crimrxiv.com/pub/c51j2rmy/release/1> (Accessed August 7, 2026).
- Angrisani, M.; Casanova, M.; Fast, N. J.; Narang, J.; and Schroeder, J. 2026. Gaps in large language model awareness, usage, and perceptions in the United States: Evidence from a nationally representative longitudinal survey. *PNAS Nexus*, 5: pgag007.
- Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Yeung, J. A.; and Pimenta, D. 2025. A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation. *NPJ Digital Medicine*, 8.
- Axon. 2026. Generate a narrative draft with Draft One. *axon.com*. <https://www.axon.com/help/draft-one/software/draft-one/draft-generation.htm> (Accessed April 30, 2026).
- Cantini, A.; and De Mauro, A. 2026. Better Prompts, Better Usefulness: A Systematic Review and Experimental Evaluation of Structured Prompting Techniques in Large Language Models. *Big Data and Cognitive Computing*, 10(7).
- Cheng, M.; Lee, C.; Khadpe, P.; Yu, S.; Han, D.; and Jurafsky, D. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391: eaec8352.
- Danry, V.; Pataranutaporn, P.; Groh, M.; and Epstein, Z. 2025. Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–31. Yokohama Japan: ACM.
- Doximity. 2026. Doximity 2026 State of AI in Medicine Report. <https://www.doximity.com/reports/state-of-ai-medicine-report/2026> (Accessed May 1, 2026).
- Eakin, D. K.; Schreiber, T. A.; and Sergeant-Marshall, S. 2003. Misinformation effects in eyewitness memory: The presence and absence of memory impairment as a function of warning and misinformation accessibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5): 813–825.
- Fanous, A.; Goldberg, J.; Agarwal, A.; Lin, J.; Zhou, A.; Xu, S.; Bikia, V.; Daneshjou, R.; and Koyejo, S. 2025. Syceval: Evaluating LLM Sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 893–900.
- Faul, F.; Erdfelder, E.; Lang, A.-G.; and Buchner, A. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2): 175–191.
- Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. <https://arxiv.org/pdf/2507.06261>.
- Goldstein, J. A.; Chao, J.; Grossman, S.; Stamos, A.; and Tomz, M. 2024. How persuasive is AI-generated propaganda? *PNAS nexus*, 3(2): pgae034.
- Greene, E.; Flynn, M. S.; and Loftus, E. F. 1982. Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior*, 21(2): 207–219.
- Gueorguieva, A.-M.; and Caliskan, A. 2026. Identifying Features Associated with Bias Against 93 Stigmatized Groups in Language Models and Guardrail Model Safety Mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 37426–37434.
- Hackenburg, K.; Tappin, B. M.; Hewitt, L.; Saunders, E.; Black, S.; Lin, H.; Fist, C.; Margetts, H.; Rand, D. G.; and Summerfield, C. 2025. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777): eaea3884.
- Highhouse, S.; and Bottrill, K. V. 1995. The influence of social (mis) information on memory for behavior in an employment interview. *Organizational Behavior and Human Decision Processes*, 62(2): 220–229.
- Kabir, S.; Udo-Imeh, D. N.; Kou, B.; and Zhang, T. 2024. Is stack overflow obsolete? An empirical study of the characteristics of chatgpt answers to stack overflow questions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.
- Karinshak, E.; Liu, S. X.; Park, J. S.; and Hancock, J. T. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–29.
- Kidd, C.; and Birhane, A. 2023. How AI can distort human beliefs. *Science*, 380(6651): 1222–1223.
- Kulkarni, P.; Liu, Y.; Fu, H.-M.; Yang, S.; Gunasekara, I.; Peloquin, M.; Spitzer-Williams, N.; Zhou, X.; Liu, X.; Ji, Z.; and Ibrahim, Y. 2025. Auto-Drafting Police Reports from Noisy ASR Outputs: A Trust-Centered LLM Approach. In *Companion Proceedings of the ACM on Web Conference 2025*, 2859–2862.
- Lee, D.; Wan, C.; Ng, P. M. L.; Fung, Y. N.; and Wu, N. 2026. Effects of AI Companions' Sycophancy and

- Emotional Mimicry on Consumers' Continuance Intention and Social Wellbeing. *International Journal of Human-Computer Interaction*, 1–23.
- Lin, H.; Czarnek, G.; Lewis, B.; White, J. P.; Berinsky, A. J.; Costello, T.; Pennycook, G.; and Rand, D. G. 2025. Persuading voters using human-artificial intelligence dialogues. *Nature*, 648(8093): 394–401.
- Lindsay, D. 1990. Misleading suggestions can impair eyewitnesses' ability to remember event details. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16: 1077–1083.
- Loftus, E. F. 1975. Leading questions and the eyewitness report. *Cognitive psychology*, 7(4): 560–572.
- Loftus, E. F. 2005. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4): 361–366.
- Loftus, E. F.; and Klemfuss, J. Z. 2024. Misinformation—past, present, and future. *Psychology, Crime & Law*, 30(4): 312–318.
- Loftus, E. F.; Miller, D. G.; and Burns, H. J. 1978. Semantic Integration of Verbal Information into a Visual Memory. *Journal of Experimental Psychology: Human Learning and Memory*, 4: 19–31.
- Loftus, E. F.; and Palmer, J. C. 1974. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5): 585–589.
- Loftus, E. F.; and Pickrell, J. E. 1995. The formation of false memories. *Psychiatric Annals*, 25(12): 720–725.
- Marvin, G.; Hellen, N.; Jjingo, D.; and Nakatumba-Nabende, J. 2024. Prompt Engineering in Large Language Models. In *International Conference on Data Intelligence and Cognitive Informatics*, 387–402. Singapore: Springer Nature Singapore.
- Milella, F.; and Cabitza, F. 2026. Perceiving AI as an Epistemic Authority or Alority: A User Study on the Human Attribution of Authority to AI. *Machine Learning and Knowledge Extraction*, 8(2): 36.
- Murphy, G.; Dawson, C. A.; Huston, C.; Ballantyne, L.; Barrett, E.; Cowman, C. S.; Fitzsimons, C.; Maher, J.; Ryan, K. M.; and Greene, C. M. 2023. Lost in the mall again: A preregistered replication and extension of Loftus & Pickrell (1995). *Memory*, 31(6): 818–830.
- Nicolas, G.; and Caliskan, A. 2025. What Are Chatbots' Stereotypes About? A Data-Driven Analysis of Large Language Models' Content Associations with Social Categories. In *Proceedings of the AAAI ACM Conference on AI, Ethics, and Society*, volume 8, 1888–1900.
- O'Donnell, R.; and Chan, J. C. 2025. Does blatantly contradictory information reduce the misinformation effect? A Registered Report replication of Loftus (1979). *Legal and Criminological Psychology*, 30(1): 3–26.
- OpenAI. 2026. Introducing GPT-5.5. <https://openai.com/index/introducing-gpt-5-5/> (Accessed August 7, 2026).
- Oukelmoun, A.; Semmar, N.; de Chalendar, G.; Cormi, C.; Oukelmoun, M.; Vibert, E.; and Allard, M. A. 2025. Detecting Omissions in LLM-Generated Medical Summaries. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 325–337.
- Pataranutaporn, P.; Archiwaranguprok, C.; Chan, S. W. T.; Loftus, E. F.; and Maes, P. 2025a. Slip Through the Chat: Subtle Injection of False Information in LLM Chatbot Conversations Increases False Memory Formation. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1297–1313. Cagliari Italy: ACM.
- Pataranutaporn, P.; Archiwaranguprok, C.; Chan, S. W. T.; Loftus, E. F.; and Maes, P. 2025b. Synthetic human memories: AI-edited images and videos can implant false memories and distort recollection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–20. Yokohama, Japan: ACM.
- Potts, J.; Billera, M.; Catren, C.; and Martinez, T. 2024. Large Language Models: Using ChatGPT for Police Leaders. <https://www.policinginstitute.org/onpolicing/large-language-models-using-chatgpt-for-police-leaders/>.
- Putnam, A. L.; Sungkhasettee, V. W.; and Roediger III, H. L. 2017. When misinformation improves memory: The effects of recollecting change. *Psychological Science*, 28(1): 36–46.
- Raj, C.; Banerjee, M.; Pan, J.; Caliskan, A.; Anastasopoulos, A.; and Zhu, Z. 2026. Talent or Luck? Evaluating Attribution Bias in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2026*, 10982–11014. San Diego, California, U.S.: Association for Computational Linguistics.
- Ramprasad, S.; Ferracane, E.; and Lipton, Z. C. 2024. Analyzing LLM Behavior in Dialogue Summarization: Unveiling Circumstantial Hallucination Trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12549–12561.
- Rawte, V.; Chakraborty, S.; Pathak, A.; Sarkar, A.; Tonmoy, S. T. I.; Chadha, A.; Sheth, A.; and Das, A. 2023. The Troubling Emergence of Hallucination in Large Language Models-An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2541–2573.
- Rockstar Games. 2013. Grand Theft Auto V [Video game]. Sony Computer Entertainment.
- Schellmann, H. 2024. *The Algorithm: How AI Decides Who Gets Hired, Monitored, Promoted, and Fired and Why We Need to Fight Back Now*. Grand Central Publishing.
- Schulhoff, S.; Ilie, M.; Balepur, N.; Kahadze, K.; Liu, A.; Si, C.; Li, Y.; Gupta, A.; Han, H.; Schulhoff, S.; Dulepet, P. S.; Vidyadhara, S.; Ki, D.; Agrawal, S.; Pham, C.; Kroiz, G.; Li, F.; Tao, H.; Srivastava, A.; Costa, H. D.; Gupta, S.; Rogers, M. L.; Goncarenco, I.; Sarli, G.; Galynter, I.; Peskoff, D.; Carpuat, M.; White, J.; Anadkat, S.; Hoyle, A.; and Resnik, P. 2025. The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. *arXiv*, arXiv:2406.06608.

- Sim, M.; Brigham, N. G.; Kohno, T.; and Caliskan, A. 2025. Biased AI Outputs Can Impact Humans' Implicit Bias: A Case Study of the Impact of Gender-Biased Text-to-Image Generators. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2375–2386.
- Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). *NIST Trustworthy and Responsible AI, National Institute of Standards and Technology*. <https://doi.org/10.6028/NIST.AI.100-1>, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225 (Accessed August 7, 2026).
- Tangermann, V. 2026. Cops Forced to Explain Why AI Generated Police Report Claimed Officer Transformed Into Frog. *Futurism*. <https://futurism.com/artificial-intelligence/ai-police-report-frog> (Accessed May 7, 2026).
- Wang, P.; Yin, K.; Tian, M.; Zheng, Y.; Wu, H.; Zhou, C.; Zhang, M.; Ma, J.; and Yuan, X. 2026. A Validation of the Human-Generative Artificial Intelligence Trust Scale. *International Journal of Human-Computer Interaction*, 42(7): 4808–4821.
- Wiechert, S.; Verschueren, A.; Ben-Shakhar, G.; Pertzov, Y.; Loftus, E. F.; and Verschuere, B. 2026. The Misinformation Effect: A Contemporary Replication and Extension of Loftus et al. (1978) to Investigate its Underlying Mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 52(7): 1225–1241.
- Wilson, K.; and Caliskan, A. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1578–1590.
- Wilson, K.; Sim, M.; Gueorguieva, A.-M.; and Caliskan, A. 2025. No Thoughts Just AI: Biased LLM Hiring Recommendations Alter Human Decision Making and Limit Human Autonomy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2692–2704.
- Xu, L.; Su, Z.; Yu, M.; Xu, J.; Choi, J. D.; Zhou, J.; and Liu, F. 2024. Identifying Factual Inconsistencies in Summaries: Grounding LLM Inference via Task Taxonomy. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14626–14641.
- Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv preprint*.
- Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A. G.; and De Choudhury, M. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. Hamburg Germany: ACM.