The Effects of Hallucination Warnings and Source Credibility on Content Learning and Source Memory when Learning with Large Language Models

Oktay Ülker[1] (oktay.uelker@uni-due.de) & Daniel Bodemer[1]
Research Methods in Psychology – Technology, Learning, Collaboration
University of Duisburg-Essen, Germany

Abstract

More and more people use large language models (LLMs) as sources of information. However, LLMs are prone to hallucinations, meaning they can produce plausible but false information. In a pre-registered laboratory experiment, we analyzed the effects of hallucination warnings (between-subjects: with vs. without) and source credibility (within-subjects: high-credible vs. unknown vs. low-credible) on content learning and source memory. N = 97 learners first received pieces of information accompanied by a source label from a simulated LLM-based chatbot, before completing learning and source memory tests. Source memory was analyzed with multinomial processing tree models. Content learning was not affected by warnings and source labels. With hallucination warnings, high- and low-credible sources were remembered better than unknown sources. However, without a warning, only low-credible sources were remembered better than unknown sources. Disclosing the source (credibility) seems promising for source memory, but especially high-credible sources may require additional highlighting in contexts without warnings.

Keywords: large language models, source credibility, hallucinations, learning, source memory, cognitive modeling

Introduction

Access to information has become easier than ever. In particular, chatbots based on large language models (LLMs, e.g., ChatGPT or DeepSeek) are increasingly used as everyday sources of information across a wide range of contexts (Chatterji et al., 2025), from general knowledge questions to decision-making and learning. Especially in educational contexts, students can use LLMs as collaborators to prepare for exams and co-construct knowledge alongside more traditional learning materials (Cress & Kimmerle, 2023; Jung & Jin, 2025). Increasingly, LLM-generated information is perceived to be as credible as information produced by humans (Huschens et al., 2023), which can increase trust in and reliance on such systems (Choung et al., 2023). However, at the same time, LLMs are prone to hallucinations: they can produce plausible-sounding but inaccurate or fabricated information (Ji et al., 2023; Sun et al., 2024). This entails the risk that students might learn and spread misconceptions or false information. In response, developers increasingly implement two design features: (1) hallucination warnings, like "*[The LLM] may produce inaccurate information about facts*." or "*[The LLM] can make mistakes. Check important info*." and (2) integrated web search with source indication (Vykopal et al., 2025; Yang, 2025). Users can then judge the reliability of information based on the credibility of its source. Some adaptations of existing LLMs integrate information regarding the quality and credibility of sources into their output, which is a feature users consider important and useful (Leiser et al., 2024).

For instance, when a student prepares for an exam and asks an LLM for information about a learning topic, the LLM can provide information and may indicate its source. The source can be a low-credible web-based forum where any user can post information without being systematically reviewed by experts. At the time of the interaction, the student can disregard the information because of low source credibility. However, when recalling the information hours or days later (e.g., during a discussion with peers without access to the LLM), the learner may remember the content information but not its (low-credible) source. Even worse, the learner could confuse the sources and incorrectly attribute the information to a high-credible source, and consequently remember unreliable information as facts.

This example highlights a central cognitive process when navigating complex information environments: *source memory*, i.e., remembering the origin of information (Johnson et al., 1993), which is critical for evaluating information credibility in retrospect or, more generally, for adaptive judgments in the future (e.g., Schaper et al., 2019). Yet it is often more fragile than content memory, especially over longer retention intervals (Kuhlmann et al., 2025). Source memory has received some attention in the context of LLM usage and recent evidence suggests that source memory in AI contexts is quite poor (Metzger et al., 2026). For example, people seem less accurate at remembering that ideas were generated with AI support than at remembering that they were generated without it (Zindulka et al., 2025).

But little is known about how users remember the source of their source, i.e., which sources LLMs relied on. This issue is particularly pronounced since LLMs may draw on high-credible but also low-credible sources, or provide information without explicit source attribution at all. Understanding how such source information is encoded and retained is therefore crucial for explaining how learners evaluate LLM-generated information over time. This leads to the central question of the present study: (how) do hallucination warnings and source labels indicating different levels of source credibility influence content learning, credibility judgments, and especially source memory?

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Hallucination Warnings

The integration of warnings regarding potential hallucinations or weaknesses (e.g., Metzger et al., 2024) has become a common feature in LLMs: users are informed that the generated content might be false. In experimental studies, contexts with and without hallucination warnings have been compared and warnings seem promising. With a warning, users seem more skeptical: they subjectively perceive more hallucinations (Hannig et al., 2025) and objectively detect more false information (Nahar et al., 2024), suggesting that warnings can foster deeper cognitive engagement with the content itself and potentially improve content learning. But do hallucination warnings also affect memory for the sources used by the LLM?

There is still limited research directly addressing how people remember sources when learning with LLMs and potential effects of hallucination warnings. It is therefore informative to draw on related areas of cognitive research to derive hypotheses. One such area is research on eyewitness testimony, where source credibility is critical and effects of warnings are tested. For example, Echterhoff and colleagues (2005) have tested the effects of different warnings on source memory: when participants were warned about potential misinformation, they monitored the sources more carefully. Source memory seems to be enhanced in contexts with more misinformation and higher skepticism (Pena et al., 2017).

In advertising psychology, where the relationship between source and information credibility is also highly relevant, contexts associated with higher skepticism (because of higher advertising exposure) are associated with better source memory (Bell et al., 2022). This finding was interpreted as a cognitive coping mechanism: in situations characterized by heightened skepticism, people allocate more attention to source information to protect themselves from potentially misleading or unreliable content. However, in such contexts with higher skepticism, memory for the content information (product statements) itself was impaired (Bell et al., 2022, Experiments 1 and 2). Together, these findings suggest that hallucination warnings may increase source memory (given learners are more skeptical), but potentially at the cost of content learning due to an item-source tradeoff (e.g., Cook et al., 2006; Jurica & Shimamura, 1999).

Source Credibility

Recipients often have to determine whether the information they receive is rather accurate or inaccurate. Decades of cognitive research have shown that one of the most important cues for evaluating information credibility is the credibility of the source itself: recipients routinely judge information as more or less reliable depending on whether its source is trustworthy or has expertise (Hovland & Weiss, 1951; Madsen & Wong, 2023; Pornpitakpan, 2004). This issue is particularly relevant in the current age of (mis)information, where online sources vary widely in credibility and artificial systems may draw on high-credible but also on low-credible sources or provide information without explicit source attribution (Yang, 2025). Given that the sources LLMs use differ in their credibility, it raises the question: does source credibility influence how well sources are remembered when learning with LLMs?

Source memory does not operate uniformly across sources. Prior work shows that sources with high or low credibility are remembered better than sources with uncertain or unknown credibility (Nadarevic & Erdfelder, 2013; Bell et al., 2022). In collaborative contexts, when receiving information from learning partners differing in their expertise, learning partners with high knowledge are remembered the best as sources (Ülker & Bodemer, 2026), especially when the content is learned for a learning test (Ülker & Bodemer, 2025). More generally, source memory seems to be enhanced when remembering a source is beneficial for future interaction or judgment (Kroneisen, 2024; Nadarevic & Erdfelder, 2019). Applied to learning with LLMs, one could hypothesize memory advantages for both high-credible sources (to identify reliable information) and low-credible sources (to avoid unreliable information), whereas remembering that a source was unknown provides less actionable guidance.

The Present Study

As LLMs are increasingly used as sources of information, hallucination warnings and source labels are sometimes used to potentially reduce the impact of misinformation. However, their (long-term) effectiveness depends on how users cognitively process such cues, particularly regarding content learning and source memory. To examine these processes, we conducted a pre-registered laboratory experiment.

Since warnings can increase engagement with the content (e.g., Nahar et al., 2024), they may also improve content learning. However, based on prior research suggesting that heightened skepticism can enhance source memory at the cost of memory for the content itself, we hypothesized that content learning is better without a hallucination warning than with a warning (*H1)* and that source memory is better with a hallucination warning than without a warning *(H2).*

Sources are often remembered better when remembering them provides a benefit. Remembering that a piece of information originated from a source with high credibility helps one judge in retrospect that this piece of information might be reliable, leading to *H3:* source memory for high-credible sources is better than source memory for unknown sources, both with (*H3a*) and without a hallucination warning (*H3b*). Correspondingly, remembering that a piece of information originated from a source with low credibility helps one judge in retrospect that this piece of information might be less reliable and potentially false, leading to *H4*: source memory for low-credible sources is better than source memory for unknown sources, both with (*H4a*) and without a hallucination warning (*H4b*).

Exploratively, we will examine how source credibility and hallucination warnings influence perceived information credibility (both while learning the information with source labels and in retrospect when sources must be remembered) and the relation between source memory and credibility judgments in retrospect.

Methods

To answer the questions raised, we conducted an experimental laboratory study with a mixed-factorial 2 × 3 design, including the between-subjects factor *hallucination warning* (with vs. without warning) and the within-subjects factor *source credibility* (high-credible vs. unknown vs. low-credible). The final sample consisted of $N = 97$ ($n_{\text{WithWarning}} = 48$, $n_{\text{WithoutWarning}} = 49$) participants (74 female, 21 male, 2 non-binary). The mean age was 21.68 years ($SD = 2.57$). Participants' prior knowledge on the learning topic (astronomy) was assessed on a subjective scale and was low on average ($M = -1.28$, $SD = 1.41$; range: −3 to +3).

Material and Procedure

(1) Introduction of the LLM and Sources After providing informed consent, participants were informed that the study aimed to test how people learn information from LLMs. They were told that they would interact with "AIsaacNewton", a (bogus) LLM-based chatbot trained on astronomical texts and (allegedly) used by physics students worldwide. In the no-warning condition, the LLM was described as trained to provide correct information. In the warning condition, participants were informed that, like other LLMs, AIsaacNewton may occasionally produce "hallucinations" (i.e., plausible-sounding but incomplete, misleading, or fabricated information) and that its output should be treated with caution.

Further, participants were informed that each statement could be accompanied by its original source. To avoid effects of prior familiarity or bias, we used two fictitious source names (AstroFacts and SpaceBase) to label sources. One label denoted a high-credible source (peer-reviewed, expert-verified scientific database, credibility rating = 9/10), and the other denoted a low-credible source (user-edited database without systematic expert review, rating = 3/10). Across participants, the assignment of these labels and names to the high- and low-credible sources was randomized, and the presentation order was individually shuffled to control for name or sequence effects. Participants were tasked to learn the provided content information on astronomy from the LLM for an upcoming learning test. However, they were not informed about the source memory test.

(2) Learning Phase Participants were asked to imagine learning with an LLM for an upcoming astronomy exam. We presented 21 standardized learning texts with similar length ($M_{\text{Words}} = 53.05$, $SD = 1.99$) on different astronomical topics.

In each trial (see Figure 1), a request for information on an astronomy topic was initiated, shown in a chat-style interface. In the condition with a warning, a red highlighted banner about possible AI hallucinations was shown first. After a short delay, the chatbot displayed content information, accompanied by a source label indicating one of three conditions: high-credible source, low-credible source, or no source label at all (7 texts each). Participants then rated the credibility of the information on a 7-point scale from −3 (*not credible at all*) to +3 (*very credible*) and could optionally type follow-up questions in a comment field (which, however, remained unanswered). Topic-source assignments were individually randomized to control for order and source effects. In each trial, participants could click "next" after 30 seconds. After 90 seconds, the page cleared automatically.

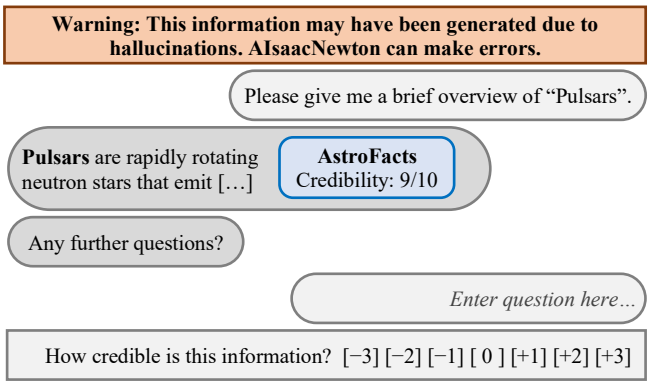

Figure 1. Simplified illustration of the learning phase.

(3) Questionnaires After the learning phase, participants answered several questions on 7-point scales from −3 (*I do not agree at all*) to +3 (*I totally agree*). Among others, we assessed whether participants were skeptical while receiving information from their learning partners ("*I was skeptical when I received information from AIsaacNewton.*") and whether the information was perceived to be hallucinated ("*I believe that some information from AIsaacNewton was hallucinated and not entirely accurate.*"). Participants also indicated how trustworthy they found both sources on scales from −3 ("*not trustworthy at all*") to +3 ("*very trustworthy*").

(4) Source Memory Test Phase Next, the source memory test followed. To follow standard procedures, the test included both old items (i.e., information presented in the learning phase) and new items (i.e., information not presented before). For each of the 21 learning topics, two old items were paraphrased from the original learning text, and one new item was created containing information relevant to the topic but novel to the participant. The source memory test thus consisted of 63 trials that all had the same structure.

Throughout the phase, the question "*Where did this information come from?*" remained on top of the screen. After 1 second, an item appeared. Then, after 1 more second, the 4 options were presented (high-credible, unknown, low-credible, new item). After a judgment, participants rated the credibility of the item on a 7-point scale (−3 to +3). After clicking "next", the page cleared and the next trial started. After 63 trials in a random order, the phase ended.

(5) Learning Test After a distractor task, participants conducted the content learning test. For each of the 21 texts in the learning phase, participants answered one multiple-choice question (one target and three distractors). Lastly, they provided demographic information and were debriefed.

Statistical Analyses

For *H1*, the dependent variable was content learning, which was operationalized as the number of correctly answered questions in the learning test (21 total). For explorative analyses, content learning was separately calculated for the information presented with different source labels (7 each).

Source memory was analyzed with multinomial processing tree (MPT) models (for an overview, see Singmann et al., 2024). The MPT model of source monitoring (Bayen et al., 1996), adapted for three sources (Keefe et al., 2002), allowed us to calculate memory processes unconfounded by guessing processes, which are often found in source memory research (e.g., Ülker & Bodemer, 2026; Zindulka et al., 2025). Our model (Figure 2) illustrates the cognitive processes involved in reconstructing the source of a piece of information. To exemplify: when a participant sees a piece of information which originated from a low-credible source (left rectangle, i = low), certain cognitive processes (parameters in the middle) can lead to certain answers in the source memory test (right rectangles). The correct answer "low" can be based on accurate memory, i.e., recognizing the item as old (D_{Low}) × remembering the source "low-credible" (d_{Low}). However, it can also be based on guessing processes, e.g., $(1 - D_{Low}) \times (1 - d_{Low}) \times a_{Low}$. The cognitive parameters are estimated based on the observed category frequencies (e.g., the number of "low" responses to items which were presented by the low-credible source). As pre-registered, to find a base model, we first restricted $D_{Unknown}$ and D_{New} in both conditions (which fitted the data) and then all D-parameters within conditions, which also fitted the data, $G^2(6) = 9.06$, $p = .170$. Our hypotheses regarding source memory were tested by restricting corresponding d-parameters (e.g., for *H2*: $d_{i\_WithWarning} = d_{i\_WithoutWarning}$, for *H3*: $d_{High} = d_{Unknown}$).

Based on effect sizes observed in Ülker and Bodemer (2026), we conducted an a priori power-analysis in multiTree (Moshagen, 2010), which revealed that given a target power of $1 - \beta = .80$, $\alpha = .05$, df = 1, and 63 items in the source memory test phase, up to 94 participants were required. We continued data collection until the end of the final survey week, therefore resulting in a total sample of more than 94 participants (final $N = 97$). The study was pre-registered (https://aspredicted.org/3m3t6m.pdf) and approved by the local ethics committee. Data can be found in the OSF (https://osf.io/fd7ge).

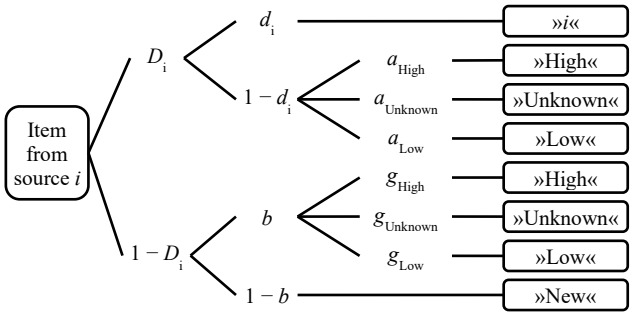

Figure 2. Simplified multinomial processing tree model.

Results

Manipulation Checks

We first checked whether the hallucination warning affected participants as intended. With a hallucination warning, participants did not perceive that the LLM hallucinated more than without a warning, $t(95) = 1.06$, $p = .290$, $d = 0.22$ ($M_{WithWarning} = 1.00$, $SD = 1.46$; $M_{WithoutWarning} = 0.69$, $SD = 1.37$). This could potentially reflect a general tendency towards assuming that—irrespective of a present warning—LLMs may hallucinate. However, since our hypotheses relied on the idea that skepticism might influence cognitive processes, we also asked participants to indicate how skeptical they were towards the information provided. Here, participants in the warning condition reported significantly higher skepticism, $t(95) = 2.94$, $p = .004$, $d = 0.60$ ($M_{WithWarning} = 1.08$, $SD = 1.23$; $M_{WithoutWarning} = 0.18$, $SD = 1.73$).

After the learning phase, participants rated the credibility of the two sources used by the LLM. Despite random assignment of information to sources, the high-credible source ($M = 1.71$, $SD = 1.17$) was judged as more credible than the low-credible source ($M = -0.32$, $SD = 1.22$), $t(96) = 13.83$, $p < .001$, $d = 1.40$, indicating a successful manipulation of source credibility through descriptions.

Content Learning (H1)

Regarding content learning (Table 1), a 2 × 3 mixed-factorial ANOVA with the between-subjects factor *hallucination warning* (with vs. without) and within-subjects factor *source credibility* (high vs. unknown vs. low) revealed no significant main effect of hallucination warning, $F(1, 95) = 2.12$, $p = .148$, $\eta_p^2 = .02$. Descriptively, and contrary to *H1*, content learning was higher with a warning. Explorative analyses showed no significant main effect of source credibility and no interaction effect, both $F(2, 190) < 0.83$, $p > .436$, $\eta_p^2 = .01$, showing that information was learned equally well irrespective of the credibility of its source.

Source Memory: Cognitive Modeling (H2, H3, H4)

We analyzed source memory (Figure 3) with MPT models to control for guessing processes. First, we tested the hypothesis that a warning should improve source memory (*H2*). While this was descriptively the case, model-based analyses revealed no significant differences, $\Delta G^2(3) = 2.16$, $p = .541$.

Further, we tested whether high-credible sources were remembered better than unknown sources (*H3*). While the difference was significant with a hallucination warning (supporting *H3a*), $\Delta G^2(1) = 9.96$, $p = .002$, the difference was not significant without a warning (contradicting *H3b*), $\Delta G^2(1) = 2.59$, $p = .108$. Testing if low-credible sources were remembered better than unknown sources (*H4*) indicated significant results in conditions with a warning (supporting *H4a*), $\Delta G^2(1) = 8.38$, $p = .004$, and without a warning (supporting *H4b*), $\Delta G^2(1) = 4.93$, $p = .026$. Explorative tests comparing memory for high- and low-credible sources were not significant in both conditions, $\Delta G^2(1) < 0.42$, $p > .518$.

Table 1: Mean content learning and credibility ratings as a function of hallucination warning and source credibility.

	Content learning		Credibility: Learning phase		Credibility: Test phase	
Source credibility	Warning+	Warning−	Warning+	Warning−	Warning+	Warning−
High	3.44 (0.24)	3.22 (0.24)	1.79 (0.13)	1.86 (0.13)	1.04 (0.11)	0.95 (0.11)
Unknown	3.58 (0.22)	2.96 (0.22)	0.53 (0.14)	0.80 (0.14)	0.93 (0.11)	0.99 (0.10)
Low	3.33 (0.24)	3.00 (0.23)	0.21 (0.14)	0.24 (0.14)	0.78 (0.10)	0.81 (0.10)
Total	10.35 (0.53)	9.18 (0.60)	0.84 (0.09)	0.97 (0.10)	0.92 (0.09)	0.92 (0.10)

Note. Warning+ = With hallucination warning, Warning− = without warning. Credibility ratings range between −3 and 3. Content learning for the single sources ranges between 0 and 7, for total between 0 and 21. Values in parentheses reflect standard errors.

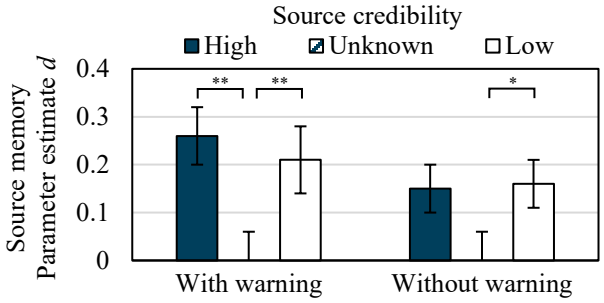

Figure 3. Source memory. *Note.* MPT model-based parameter estimates *d* reflect the probability of remembering the source of a piece of information (range: 0 – 1). Error bars reflect standard errors. $^{*}p < .05$, $^{**}p < .01$

Information Credibility (Explorative)

We also explored whether perceived *information credibility* (Table 1) differed as a function of hallucination warning and source credibility in two phases: (1) in the learning phase (with available source information) and (2) in the test phase (when source information had to be remembered).

In the learning phase, source credibility had a large significant effect on information credibility, $F(2, 190) = 98.58$, $p < .001$, $\eta_p^2 = .51$: information from high-credible sources was perceived as most credible, followed by unknown sources, and lastly low-credible sources, all $t(95) > 3.89$, $p_{Holm} < .001$. This indicates that participants attended source labels, which is a prerequisite for further (source) memory processes. We found no main effect of hallucination warning, $F(1, 95) = 0.76$, $p = .385$, $\eta_p^2 = .01$, and no interaction effect, $F(2, 190) = 0.64$, $p = .529$, $\eta_p^2 = .01$.

In the test phase, the effect of (original) source credibility on information credibility persisted but was smaller, $F(2, 190) = 7.65$, $p < .001$, $\eta_p^2 = .07$, possibly reflecting decay of source information in memory. Information from high-credible sources was rated as more credible than information from low-credible sources, $t(95) = 3.84$, $p_{Holm} = .001$, but not more credible than information from unknown sources, $t(95) = 0.71$, $p_{Holm} = .481$. Information from unknown sources was rated as more credible than information from low-credible sources, $t(95) = 2.75$, $p_{Holm} = .014$. Again, we found no main effect of hallucination warning, $F(1, 95) < 0.01$, $p = .996$, $\eta_p^2 < .01$, and no interaction effect, $F(2, 190) = 0.97$, $p = .381$, $\eta_p^2 = .01$.

Relationships Between Variables (Explorative)

We also explored relationships between several variables. For source memory, our MPT-approach provides probability estimates at the group level and thus does not allow testing correlations at the individual level. Therefore, we used conditional source identification measures, which reflect the probability of correctly identifying a source given correct item recognition. However, such measures can confound memory and guessing (Bröder & Meiser, 2007). Indeed, we found significant guessing biases: when sources could not be remembered, high-credible sources and unknown sources were over-guessed, $\Delta G^2(2) > 7.82$, $p < .020$. In line with our pre-registered approach, we did not analyze individual source identification measures but instead calculated the average across all three conditional source identification scores ($M_{WithWarning} = .40$, $SD = 0.09$, $M_{WithoutWarning} = .38$, $SD = 0.09$), which we used as a proxy for source memory.

First, we found a positive and significant relationship between source memory and content learning, $r(95) = .29$, $p = .004$. However, there was no relationship between source memory and skepticism, $r(95) = -.01$, $p = .892$, or perceived hallucinations, $r(95) = .16$, $p = .123$. Furthermore, we tested whether source memory predicted credibility ratings in the test phase to examine the idea that accurate source identification supports maintaining credibility assessments in retrospect after the source label is no longer present. Overall model tests of the regression models can be found in Table 2.

Table 2. Linear regression models for source memory (predictor) and credibility ratings in the test phase (outcome).

Source credibility	Overall model test
High	$F(1, 95) = 7.24$, $p = .008$, $R^2 = .07$
Unknown	$F(1, 95) = 0.04$, $p = .833$, $R^2 < .01$
Low	$F(1, 95) = 0.43$, $p = .512$, $R^2 < .01$

Source memory significantly predicted credibility ratings for information originally from high-credible sources ($b = 2.20$, $\beta = .27$): better source memory was associated with higher credibility ratings. However, there were no relationships with credibility ratings for items from unknown sources ($b = 0.17$, $\beta = .02$) or low-credible sources ($b = -0.52$, $\beta = -.07$). This supports the notion that accurate source memory is relevant for credibility judgments in retrospect, but only for information originally from high-credible sources.

Discussion

In an era of rapidly advancing AI, accessing information via large language models (LLMs) has become effortless. However, LLMs are prone to hallucinations (that is, plausible but false information). In this pre-registered study, we examined how hallucination warnings (with vs. without) and source credibility (high vs. unknown vs. low) affect content learning and memory for the sources used by the LLM.

Contrary to our hypothesis, hallucination warnings did not impair content learning, and performance was even slightly higher in the warning condition. At this point, it is important to note that our first hypothesis was derived from source memory experiments without explicit learning goals (e.g., Bell et al., 2022). In contrast, in the present study, there was a clear task to learn informational texts, which puts the focus on the content while still providing salient source labels. For instance, based on learning theories like the ICAP framework (Chi & Wylie, 2014), learning is increased with higher (cognitive) engagement and deeper elaboration. Hallucination warnings may draw attention to the content and prompt more elaborative processing of the material, which can be associated with content learning. In addition, the positive relationship between content learning and source memory in the present study aligns with findings from instructional contexts where attention to content and contextual features (such as sources) can enhance each other (Bodemer & Dehler, 2011; Ülker & Bodemer, 2025).

Surprisingly, the presence of a hallucination warning did not significantly improve source memory overall (potentially because participants were wary of hallucinations, irrespective of a warning). Combined with the missing relationship between source memory and reported skepticism, this suggests that the driving factor for changes in source memory is not skepticism (see also Nadarevic & Bell, 2024). However, source memory for specific sources differed. Without a warning, we only found a memory advantage for low-credible sources over unknown sources, but not for high-credible sources. In contrast, with a warning, both low- and (especially) high-credible sources were better remembered than unknown sources. Warnings may shift attentional goals and render high-credible sources important, making them more salient and memorable, possibly because learners interpret warnings as a cue to focus on determining which sources can be trusted at all. A potential explanation could also lie within the expectation-violation model: source memory tends to be better for less expected sources, both in short-term (e.g., Zheng et al., 2025) and long-term (e.g., Ehrenberg & Klauer, 2005). Hallucination warnings might prime that some information is unreliable, potentially making high-credible sources less expected and thereby somewhat heightening memory for such sources.

Furthermore, when learners had to judge the credibility of information in retrospect with absent source labels, better source memory predicted higher credibility ratings for information originally from high-credible sources, which supports the functional nature of source memory and its importance for adaptive decisions.

Methodologically, our pseudo-interactive context using fictional sources and the bogus AI system avoided bias from prior experiences with real sources and a real AI system. However, such prior experiences can shape mental models about AI partners in real-world contexts (Gessinger et al., 2025), which could affect source memory. We also deliberately avoided truly hallucinated material: (1) for ethical reasons, since even after debriefing, humans might forget that a piece of information was labeled as "false" (Nadarevic & Erdfelder, 2019; Nadarevic, 2025), and (2) to keep information constant across conditions.

Of course, this came at the cost of limited external validity. For example, in our study, the time between learning and testing was only a few minutes (potentially inflating source memory), whereas in real contexts, content and source information are often recalled after longer intervals. Although overall source memory declines over time, the same credibility-related patterns can persist even after delays (Buchner et al., 2009; Nadarevic & Erdfelder, 2013), suggesting that the patterns observed in our test may generalize to longer retention intervals. Furthermore, while our sample had rather low prior knowledge of the learning topic, in different contexts, users with high prior knowledge may also use LLMs to learn. Higher prior knowledge is sometimes associated with better overall source memory (e.g., Stang Lund et al., 2019). However, effects of prior knowledge on source memory should be tested with objective multiple-choice tests assessing prior knowledge instead of subjective self-ratings, which were used in our study (Anmarkrud et al., 2022).

Outlook and Conclusion

Practically, while hallucination warnings and source labels seem promising, they also need refinement. (1) By making source labels salient through visual cues and explicit credibility ratings, learners can use them not only during interaction (Leiser et al., 2024) but also to form durable source-content links in long-term memory. Our findings indicate that this seems to be particularly important for high-credible sources in contexts without warnings. (2) Given that pre-warnings did not improve source memory per se, another approach could be post-warnings delivered after information presentation, which can enhance source memory (Echterhoff et al., 2005). (3) Prompts (e.g., Pennycook & Rand, 2022) could accompany general hallucination warnings: actionable prompts that encourage active engagement and (source) verification processes could potentially promote deeper processing of sources through engagement.

In today's diverse information environments, source memory plays a crucial role in making accurate credibility judgments in retrospect. Extending work on hallucination warnings or source qualifiers (e.g., Leiser et al., 2024; Nahar et al., 2024), our study demonstrated that such cues can be used to build durable source-content links in long-term memory. Ultimately, considering design implications for source memory in future applications can help users to effectively utilize LLMs in acquiring (reliable) knowledge.

Acknowledgments

The authors thank Thu Le for her work in creating the study environment and collecting the data in the course of her Bachelor's Thesis.

References

Anmarkrud, Ø., Bråten, I., Florit, E., & Mason, L. (2022). The role of individual differences in sourcing: A systematic review. *Educational Psychology Review*, *34*(2), 749–792. https://doi.org/10.1007/s10648-021-09640-7

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 197–215. https://doi.org/10.1037/0278-7393.22.1.197

Bell, R., Mieth, L., & Buchner, A. (2022). Coping with high advertising exposure: A source-monitoring perspective. *Cognitive Research: Principles and Implications*, *7*(1), 82. https://doi.org/10.1186/s41235-022-00433-2

Bodemer, D., & Dehler, J. (2011). Group awareness in CSCL environments. *Computers in Human Behavior, 27*(3), 1043–1045. https://doi.org/10.1016/j.chb.2010.07.014

Bröder, A., & Meiser, T. (2007). Measuring source memory. *Zeitschrift für Psychologie/Journal of Psychology, 215*(1), 52–60. https://doi.org/10.1027/0044-3409.215.1.52

Buchner, A., Bell, R., Mehl, B., & Musch, J. (2009). No enhanced recognition memory, but better source memory for faces of cheaters. *Evolution and Human Behavior*, *30*(3), 212–224. https://doi.org/10.1016/j.evolhumbehav.2009.01.004

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How people use ChatGPT* (Report No. w34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction, 39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Cook, G. I., Marsh, R. L., & Hicks, J. L. (2006). Source memory in the absence of successful cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 828–835. https://doi.org/10.1037/0278-7393.32.4.828

Cress, U., & Kimmerle, J. (2023). Co-constructing knowledge with generative AI tools: Reflections from a CSCL perspective. *International Journal of Computer-Supported Collaborative Learning, 18*(4), 607–614. https://doi.org/10.1007/s11412-023-09409-w

Echterhoff, G., Hirst, W., & Hussy, W. (2005). How eyewitnesses resist misinformation: Social postwarnings and the monitoring of memory characteristics. *Memory & Cognition*, *33*(5), 770–782. https://doi.org/10.3758/BF03193073

Ehrenberg, K., & Klauer, K. C. (2005). Flexible use of source information: Processing components of the inconsistency effect in person memory. *Journal of Experimental Social Psychology, 41*(4), 369–387. https://doi.org/10.1016/j.jesp.2004.08.001

Gessinger, I., Seaborn, K., Steeds, M., & Cowan, B. R. (2025). ChatGPT and me: First-time and experienced users' perceptions of ChatGPT's communicative ability as a dialogue partner. *International Journal of Human-Computer Studies, 194*, 103400. https://doi.org/10.1016/j.ijhcs.2024.103400

Hannig, L., Bush, A., Aksoy, M., Trappen, T., Becker, S., & Ontrup, G. (2025). *Campus AI vs. commercial AI: How customizations shape trust and usage of LLM as-a-service chatbots* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2509.15826

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15*(4), 635–650. https://doi.org/10.1086/266350

Huschens, M., Briesch, M., Sobania, D., & Rothlauf, F. (2023). *Do you trust ChatGPT? – Perceived credibility of human and AI-generated content* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.02524

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Johnson, M. K., Hashtroudi, S., & Lindsay, D. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Jung, Y., & Jin, S. H. (2025). Questioning the role of AI as collaborator: A systematic literature review of generative AI-supported knowledge construction. *Interactive Learning Environments,* 1–20. https://doi.org/10.1080/10494820.2025.2556808

Jurica, P. J., & Shimamura, A. P. (1999). Monitoring item and source information: Evidence for a negative generation effect in source memory. *Memory & Cognition, 27*(4), 648–656. https://doi.org/10.3758/BF03211558

Keefe, R. S. E., Arnold, M. C., Bayen, U. J., McEvoy, J. P., & Wilson, W. H. (2002). Source-monitoring deficits for self-generated stimuli in schizophrenia: Multinomial modeling of data from three sources. *Schizophrenia Research, 57*(1), 51–68. https://doi.org/10.1016/S0920-9964(01)00306-1

Kroneisen, M. (2024). Context matters: The influence of the social situation on source memory. *Culture and Evolution*, *20*(1), 111–123. https://doi.org/10.1556/2055.2023.00039

Kuhlmann, B. G., Symeonidou, N., Tanyas, H., Mitchell, K. J., & Naveh-Benjamin, M. (2025). *Dissociated forgetting patterns in item versus source memory* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fh68e_v1

Leiser, F., Eckhardt, S., Leuthe, V., Knaeble, M., Maedche, A., Schwabe, G., & Sunyaev, A. (2024). Hill: A hallucination identifier for large language models. In

Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Article 482 (pp. 1–13). https://doi.org/10.1145/3613904.3642428

Madsen, J. K., & Wong, M. (2023). Comparing predictions from the Elaboration Likelihood Model and a Bayesian model of argumentation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45. Retrieved from https://escholarship.org/uc/item/75x869j9

Metzger, L., Bühler, N., Geiger, S., & Erdfelder, E. (2026). Who—or what—told you? Source memory for AI- vs. human-authored content. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26)*, Article 777 (pp. 1–5). https://doi.org/10.1145/3772363.3799350

Metzger, L., Miller, L., Baumann, M., & Kraus, J. (2024). Empowering calibrated (dis-)trust in conversational agents: A user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Article 481 (pp. 1–19). https://doi.org/10.1145/3613904.3642122

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*(1), 42–54. https://doi.org/10.3758/BRM.42.1.42

Nadarevic, L. (2025). *Effects of statement type and study context on memory for truth and falsity* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9t8fb_v3

Nadarevic, L., & Bell, R. (2024). Remembering the truth or falsity of advertising claims: A preregistered model-based test of three competing theoretical accounts. *Psychonomic Bulletin & Review*, *31*(5), 2323–2331. https://doi.org/10.3758/s13423-024-02482-8

Nadarevic, L., & Erdfelder, E. (2013). Spinoza's error: Memory for truth and falsity. *Memory & Cognition, 41*(2), 176–186. https://doi.org/10.3758/s13421-012-0251-z

Nadarevic, L., & Erdfelder, E. (2019). More evidence against the Spinozan model: Cognitive load diminishes memory for "true" feedback. *Memory & Cognition*, *47*, 1386–1400. https://doi.org/10.3758/s13421-019-00940-6

Nahar, M., Seo, H., Lee, E. J., Xiong, A., & Lee, D. (2024). *Fakes of varying shades: How warning affects human perception and engagement regarding LLM hallucinations* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2404.03745

Pena, M. M., Klemfuss, J. Z., Loftus, E. F., & Mindthoff, A. (2017). The effects of exposure to differing amounts of misinformation and source credibility perception on source monitoring and memory accuracy. *Psychology of Consciousness: Theory, Research, and Practice, 4*(4), 337–347. https://doi.org/10.1037/cns0000137

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications, 13*, 2333. https://doi.org/10.1038/s41467-022-30073-5

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Schaper, M. L., Mieth, L., & Bell, R. (2019). Adaptive memory: Source memory is positively associated with adaptive social decision making. *Cognition*, *186*, 7–14. https://doi.org/10.1016/j.cognition.2019.01.014

Singmann, H., Heck, D. W., Barth, M., Erdfelder, E., Arnold, N. R., Aust, F., Calanchini, J., Gümüsdagli, F. E., Horn, S. S., Kellen, D., Klauer, K. C., Matzke, D., Meissner, F., Michalkiewicz, M., Schaper, M. L., Stahl, C., Kuhlmann, B. G., & Groß, J. (2024). Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic review of multinomial processing tree models across the multiverse of estimation methods. *Psychological Bulletin, 150*(8), 965–1003. https://doi.org/10.1037/bul0000434

Stang Lund, E., Bråten, I., Brandmo, C., Brante, E. W., & Strømsø, H. I. (2019). Direct and indirect effects of textual and individual factors on source-content integration when reading about a socio-scientific issue. *Reading and Writing*, *32*(2), 335–356. https://doi.org/10.1007/s11145-018-9868-z

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications, 11,* 1–14. https://doi.org/10.1057/s41599-024-03811-x

Ülker, O., & Bodemer, D. (2025). "Memorize the source!" — Does focusing on learning partners improve source memory without impairing content learning?. In Oshima, J., Chen, B., Vogel, F., & Järvelä, J. (Eds.), *Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning – CSCL 2025* (pp. 90–98). International Society of the Learning Sciences. https://doi.org/10.22318/cscl2025.479268

Ülker, O., & Bodemer, D. (2026). Source memory and collaborative learning: The roles of conflicting information, group composition, and learning partner expertise. *Frontiers in Education*, *10*, 1691038. https://doi.org/10.3389/feduc.2025.1691038

Vykopal, I., Pikuliak, M., Ostermann, S., & Šimko, M. (2025). *Assessing web search credibility and response groundedness in chat assistants* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.13749

Yang, K. C. (2025). *News source citing patterns in AI search systems* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.05301

Zheng, J., Yu, J., Xu, M., Guan, C., Fu, Y., Shen, M., & Chen, H. (2025). Expectation violation enhances short-term source memory. *Psychonomic Bulletin & Review, 32*(6), 2893–2902. https://doi.org/10.3758/s13423-025-02715-4

Zindulka, T., Goller, S., Fernandes, D., Welsche, R., & Buschek, D. (2025). *The AI memory gap: Users misremember what they created with AI or without* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2509.11851