Overconfidence without Understanding? Evidence for an Illusion of Understanding in AI-Assisted Explanations

Ivan Aslanov[1] (ivaslanov@gmail.com), Patricio Felmer[1], & Ernesto Guerra[1]
[1]Center for Advanced Research in Education, Universidad de Chile

Abstract

The Illusion of Explanatory Depth (IOED), i.e. the tendency to overestimate one's understanding, has been shown to increase when people search for explanations online. However, it remains unclear whether this effect extends to explanations obtained from AI chatbots. We examined whether chatbot use increases IOED and how it affects explanatory quality. In a between-subjects experiment, one group consulted ChatGPT on four topics, a second group read texts identical to the chatbot's responses without knowing their source, and a control group received no materials. Participants then rated their ability to explain each topic, produced written explanations, and evaluated their explanations. Results showed that the AI group exhibited the greatest overestimation of explanatory ability. In addition, both coder ratings and automated text analyses indicated lower explanation quality in the AI group compared to the non-AI group.

Keywords: explanations; Illusion of Explanatory Depth; AI; ChatGPT

Introduction

The explosive growth of AI assistants based on LLMs has sparked debates about their possible benefits and risks on users' literacy and reasoning skills (Guerra et al., 2025; Huettig & Christiansen, 2024). Considering that more and more people are using AI assistants to search for and explain complex, including scientific, information (Greussing et al., 2025), an important question remains: to what extent the explanations obtained from AI support real understanding? As Messeri & Crockett (2024) note, AI technologies promise to enhance productivity and objectivity, but they can also exploit our heuristics, evoking an illusion of understanding. In the classic work of Rozenblit & Keil (2002), the Illusion of Explanatory Depth (IOED) is described as overestimation of how coherent and detailed people's understanding is. This bias is deeply rooted in our cognitive system, manifesting at least as early as elementary school age (Mills & Keil, 2004). The strength of IOED depends on cognitive characteristics of participants, such as adopting a concrete or abstract construal style (Alter et al., 2010) or on the executive resources available to assess understanding (Gaviria & Corredor, 2025).

On the other hand, it may depend on the knowledge domain and have a partly social nature, prompting experts to overestimate their explanatory knowledge in their field of expertise (Fisher & Keil, 2016) or leading people to overstate their knowledge in socially desirable topics (Gaviria & Corredor, 2021).

However, the context in which explanatory information is obtained may also be important: it has been shown that IOED can be amplified when individuals seek explanations online (Fisher et al., 2015). This leads to overconfidence in one's knowledge (Ward, 2021) and, at the same time, reduces the amount of information retained in memory compared to cases where participants read identical texts directly, bypassing the internet search stage (Fisher et al., 2022). Thus, easier access to online information may reduce the amount of information retained in memory. Fisher and Oppenheimer (2021) argue that the overestimation of one's own understanding is driven by a misattribution of knowledge, similar to what occurs in collaborative problem solving within teams. In contrast, Eliseev and Marsh (2023), replicating the same effect, attribute its source to the interface of search engines. Hence, testing this illusion in novel contexts in which people interact with AI chatbots (contexts that differ from classical Internet search both in the format of knowledge acquisition and in the interface) represents a relevant research question.

This report describes a study aimed at testing whether an increase in the illusion of understanding emerges during users' interactions with chatbots. The study employs a IOED assessment procedure, as well as two independent methods for evaluating the quality of the explanations produced by participants. The research design was not intended to test hypotheses about the underlying cognitive mechanisms and was focused solely on establishing the presence of the effect; however, potential mechanisms are briefly discussed in the Discussion section in the context of subsequent experiments. The data analysis rationale and the results for each method are described separately in the corresponding subsections.

Method

Sample size rationale

Studies examining the magnitude of the IOED at the between-subjects level mostly include specific cognitive functions and personal traits as predictors, which are not part of the present study design (see Gaviria & Corredor, 2021;

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

Gaviria & Corredor, 2025; Körner et al., 2024). On the other hand, studies investigating the emergence of the IOED in the context of Internet search do not assess the magnitude of the IOED and instead focus on differences in pre-ratings across conditions (see Fisher et al., 2015; Eliseev & Marsh, 2023). For this reason, running a simulation for sample size estimation would be not entirely reliable, since the literature lacks studies describing model results that could be approximated in our context. Therefore, we relied on the calculation by Körner et al. (2024), who, after reviewing a set of previously published IOED experiments, concluded that for studying between-subjects effects in the IOED paradigm, a total of 62 participants, or about 31 per group (for two groups), is sufficient for power 0.8. Since our experiment involves three groups, we increased the target sample size to 34 participants per group.

Participants

We recruited university students by advertising our study via social media and at universities' news websites. With the aim at achieving a balanced design with 34 participants in each of the three groups we collected 102 responses. After reviewing them we had to exclude 10 participants who provided at least one blank answer as well as those who didn't use a custom version of ChatGPT designed for this study. Then, 10 more were invited to maintain the equal number of participants across the groups, reaching the final sample with 102 participants (M age = 21.3, women = 36). Every participant received a payment of approximately 2 USD for participation and signed the informed consent. Moreover, participants reported the frequency of AI-bots´ usage in their everyday lives: 31.4% used them several times a day; 26.5% - almost every day; 25.5% - approximately once a week or more, but not every day; 14.7% - more than once a month but less than once a week; 1.0% - once a month or less frequently; 1.0% reported not using them at all.

Materials, design, and procedure

The experiment was developed using lab.js and conducted online on the open-lab.online platform. We relied on the procedure presented by Eliseev and Marsh (2023) to measure how Internet search enhances individuals' evaluation of their own understanding. However, we slightly modified the procedure by additionally including a phase in which participants evaluated the quality of their own explanations, as our focus was on interaction with ChatGPT and we were specifically interested in the difference between pre- and post-explanation ratings rather than in the initial differences in pre-ratings between groups (see Figure 1). Participants were randomly assigned to one of three groups: GPT, no-GPT, and control. In each group, participants were asked the same four questions (presented randomly) that required general knowledge to provide an explanation selected from Eliseev & Marsh (2023):

1. Why are there leap years?
2. How is glass made?
3. How do zippers work?
4. Why are there phases of the moon?

The main task for participants consisted in three steps: i) to predict how well they could explain each question, ii) to write an answer for each question and, iii) to rate the quality of their written explanations (in all cases, a 7-point scale was used, 1 – very poor, 7 – very well). However, prior to completing the main task, the GPT and no-GPT groups underwent an induction phase. Participants in the GPT group posed each of the questions to ChatGPT. We used a customized version of ChatGPT that provided identical responses to all users (the responses were pre-generated using ChatGPT-4.5). According to the experimental protocol, participants asked the questions in a predefined format and were not allowed to ask follow-up questions. This procedure ensured that all participants received the same amount of information and had equivalent interaction experiences with the chatbot (we acknowledge that this approach reduces ecological validity; however, within the scope of this study, our primary aim was to control for the equivalence of the information provided.)

In contrast, the no-GPT group simply read the same texts that participants in the GPT group received from the chatbot. These texts were presented as supplementary materials, and participants were not informed about their source. After this, participants completed the main task, as described above.

We should also discuss one theoretical and methodological aspect of this study. The procedure described in Eliseev & Marsh (2023), which served as the basis for the present experiment, differs from the procedure used in the classic study by Rozenblit & Keil (2002). Specifically, in Eliseev and Marsh, the key question concerns participants' *ability to explain* (i.e., "how well can you explain…"), whereas in Rozenblit and Keil the question focuses on *depth of understanding*. At the same time, in another study by Fisher, Goddu, and Keil (2015), in which the authors examined the impact of the Internet on evaluations of explanatory knowledge, participants were asked to rate their *ability to explain the answer to a question* (although the term IOED is not explicitly used in that work). In light of these differences, we considered it important to explicitly acknowledge them here.

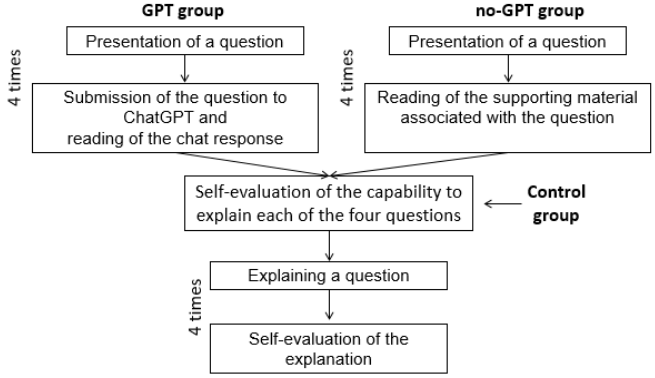

Figure 1: Experimental procedure

Data analysis

This study employed three distinct analytical approaches. Specifically, we analyzed: (1) participants' self-assessments provided before and after they produced their explanations; (2) quality ratings assigned by independent coders after reading participants' explanations; and (3) results obtained from the automated text analysis of the explanations. Below, we describe in greater detail each of these approaches.

Participants' self-assessments

For each item, two ratings were obtained: a pre-rating (self-evaluation of how well the participant could explain the phenomenon) and a post-rating (self-evaluation of the explanation they produced).

We fitted a linear mixed-effects model using the *lme4* package in R (R Core Team, 2023). The dependent variable was the self-rating, with condition (GPT, no-GPT, control), time (pre vs. post), and their interaction as fixed effects. The GPT condition and the pre-rating were set as the intercept. We implemented a maximal random-effects structure, including random intercepts and random slopes for time by both participants and items. To assess model fit, we compared this specification to several reduced models, specifically a model without random slopes for items, a model without random slopes for participants, and a model with random intercepts only. Likelihood ratio tests (*anova* function in R) showed no significant differences in model fit between the maximal specification and the reduced models. Thus, we maintained the maximal random-effects structure justified by the design (Barr et al., 2013). We obtained p-values with the *lmerTest* package (Kuznetsova et al., 2017).

Coder-assigned quality ratings

To obtain more objective evaluations of the explanations across the three conditions, we recruited two independent coders. They were asked to rate each explanation using a four-point scale: 1 – the response was completely incorrect; 2 – the general idea was described correctly, but important details were missing; 3 – the general idea was described correctly and included some details, but the response still contained logical gaps or clear factual errors; 4 – the general idea was correctly described, the response was logically complete, and no clear factual errors were present.

Prior to coding, each coder read the stimulus materials from the induction phase as well as an additional explanation for each question taken from reliable sources (e.g., National Geographic website). This was done to ensure that coders were not restricted solely to the information contained in the stimuli but had more diverse prior knowledge.

The coders then independently evaluated approximately 9% of the explanation corpus, noting aspects of the texts they considered important for assessment. This subset of 9% included the three shortest, three longest, and three "average-length" explanations for each question (36 in total), in order to expose coders to a variety of texts during training. At this stage, inter-rater reliability was high, with Cohen's Kappa using squared weights indicating strong agreement, $\kappa = .86$, $z(36) = 5.35$, $p < .001$. Afterwards, the coders discussed the aspects of the texts they considered important as well as all differences in their ratings, and reached consensus on final ratings and criteria for evaluating the texts. Based on this process, they proceeded to rate the remaining texts in the corpus. Final inter-rater reliability was again high, squared Cohen's $\kappa = .82$, $z(408) = 16.70$, $p < .001$. The coders then discussed any residual disagreements and produced the final set of ratings. Importantly, the coders did not know the experimental hypothesis and the number of experimental groups.

Finally, to analyze the ordinal ratings provided by the coders, we fitted a cumulative link mixed model (CLMM) with a logit link and flexible thresholds using the Laplace approximation, implemented in R with the *ordinal* package (Christensen, 2023). The model included condition as a fixed effect and random intercepts for both participants and items.

Automated text analysis

All textual responses were preprocessed in R: text was lowercased, contractions expanded with the *textclean* package (Rinker, 2018), punctuation removed, whitespace normalized, and Spanish stop words removed using the Snowball stemmer list (Porter, 2001). Each response received a unique identifier for traceability.

We then calculated three dependent variables: response length, lexical diversity, and semantic similarity to the prompt (the stimuli from the induction phase). Response length was measured as word count. Lexical diversity was assessed with the Corrected Type-Token Ratio (CTTR, Carroll, 1964), a measure less sensitive to text length, computed via *quanteda.textstats* package (Benoit et al., 2018). Semantic similarity was assessed using a TF-IDF vector space model (Salton & McGill, 1983) implemented in *text2vec* package (Selivanov et al., 2020). Prompts were preprocessed identically, a joint vocabulary was built, and responses and prompts were transformed into TF-IDF weighted document-term matrices. Semantic similarity was quantified with cosine similarity (Turney & Pantel, 2010), ranging from 0 (no overlap) to 1 (identical).

To analyze the effect of the experimental conditions we implemented linear mixed-effects regression models using REML estimation. In all three models we set the GPT condition as intercept, such as that we could contrast that condition against the no-GPT and control conditions. Models included random intercept for participants and items.

Results

Participants' self-assessments

The mixed-effects regression analysis (Table 1) showed that, relative to the GPT condition, the control group reported significantly lower baseline ratings ($\beta = –1.44$, $SE = 0.26$, $t = –5.56$, $p < .001$), whereas the no-GPT group did not differ significantly ($\beta = –0.14$, $SE = 0.26$, $t = –0.54$, $p = .59$).

Across conditions, post-test ratings were overall lower than pre-test ratings (β = –1.32, *SE* = 0.20, *t* = –6.56, *p* < .001).

Table 1: Mixed-effect regression analysis outcomes for the effects of condition and time on explanatory ratings

	Estimates	SE	t	p
(Intercept)	5.42	0.32	17.05	<0.001***
condition [no-GPT]	-0.14	0.26	-0.54	0.590
condition [control]	-1.44	0.26	-5.56	<0.001***
time [post]	-1.32	0.20	-6.56	<0.001***
condition [no-GPT] × time [post]	0.54	0.25	2.20	0.028*
condition [control] × time [post]	0.89	0.25	3.60	<0.001***

* p<0.05, *** p<0.001

Crucially, the interaction between condition and time indicated that the reduction from pre- to post-test in the GPT condition was significantly smaller in the no-GPT condition, β = 0.54, *SE* = 0.25, *t* = 2.20, *p* = .028, and even more so in the control condition, β = 0.89, *SE* = 0.25, *t* = 3.60, *p* < .001. This suggests that the difference in IOED between GPT and no-GPT conditions was reliable, supporting our main hypothesis (see Figure 2).

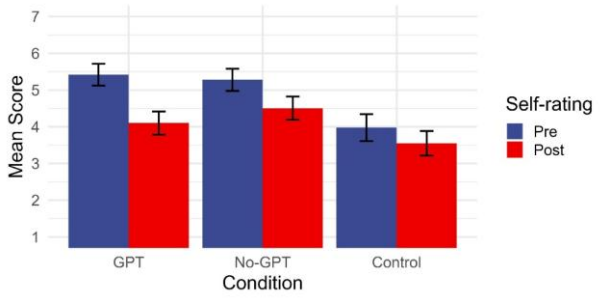

Figure 2: Mean self-ratings across three conditions before (pre-) and after (post-) giving an explanation

Coder-assigned quality ratings

The mean scores were 2.47 (*SD* = 0.91) for the GPT group, 2.75 (*SD* = 0.89) for the No-GPT group, and 1.8 (*SD* = 0.71) for the Control group. The results of CLMM showed that, relative to the GPT condition, ratings in the no-GPT condition were significantly higher, β = 0.83, *SE* = 0.41, *z* = 2.01, *p* = .044, whereas ratings in the control condition were significantly lower, *b* = –1.86, *SE* = 0.43, *z* = –4.31, *p* < .001. Thus, explanations in the no-GPT group had somewhat higher odds of receiving better ratings compared to the GPT group, whereas the control group, as expected, showed the lowest results.

Automated text analysis

As shown in Table 2, the mixed-effects models revealed significant differences across conditions for all linguistic measures. In the GPT condition, responses averaged 43.82 words. The no-GPT group produced significantly longer responses (+15.82 words; ~59.64), while the control group showed a non-significant trend toward shorter responses (–10.27 words; ~33.55). Lexical diversity followed a similar pattern: the GPT condition had a baseline CTTR of 2.71, the no-GPT group scored higher (+0.29; ~3.0), and the control group lower (–0.28; ~2.43). Semantic similarity was 0.22 in the GPT condition, higher in the no-GPT group (+0.03; ~0.25), and significantly lower in the control group (–0.06; ~0.16).

Table 2: Mixed effect regression analysis outcomes for all three dependent variables; number of words (Word count), lexical diversity (CTTR) and semantic similarity (Cosine distance) between response and prompts

		Est.	Se	t	p
Word count	Intercept	43.82	7.38	5.94	0.001**
	No-GPT	15.82	5.82	2.72	0.008**
	Control	-10.27	5.82	-1.76	0.081
		Est.	se	t	p
CTTR	Intercept	2.71	0.12	22.38	<0.001***
	No-GPT	0.29	0.13	2.16	0.033*
	Control	-0.28	0.13	-2.07	0.041*
		Est.	se	t	p
Cosine distance	Intercept	0.22	0.04	5.53	0.009**
	No-GPT	0.03	0.01	2.41	0.018*
	Control	-0.06	0.01	-4.67	<0.001***

* p<0.05 ** p<0.01 *** p<0.001

Overall, the no-GPT condition consistently exhibited higher values than the GPT condition across all measures; this condition produced longer responses with richer vocabulary and higher semantic fidelity to the prompts. The control condition generally showed lower values in all measures relative to both GPT-assisted conditions, particularly in semantic alignment with the original prompts. This last effect is highly expected given that in the control condition, participants did not see any prompts.

Discussion

Modern AI bots generate texts that imitate reasoning and explanations. The present study investigates the extent to which such explanations can provoke IOED when people rely on them. We employed three different metrics to examine this process. First, we compared metacognitive predictions of one's explanatory knowledge with the evaluation of actual explanations. We found that the difference between these two measures was greatest in the GPT condition. Importantly, this cannot be explained solely by the linguistic properties of the stimulus material, since the GPT and no-GPT groups were presented with identical content. Among all three groups, predictions of the quality of one's explanations were most misaligned with actual evaluations in the group that interacted with ChatGPT.

At the same time, the tendency to overestimate one's own explanatory abilities in this group appears to be linked to genuinely lower retention of explanatory knowledge in memory. Participants not only believed that their explanations were weaker than they predicted, but also provided less coherent and substantive explanations on average, as confirmed by coders. It is quite expected that among the three groups, the lowest score was obtained by the control group, in which participants had no supporting materials. On the other hand, it is important that the no-GPT group obtained higher ratings, highlighting the presence of a greater number of details and more complete causal connections.

Linguistic analyses of the texts also support this interpretation. Explanations in the GPT group were shorter and less lexically diverse than in the no-GPT group; on the other hand, they were less similar to the stimulus texts, suggesting that participants remembered and reproduced explanations less effectively when receiving them from the chatbot.

Taken together, these results suggest that obtaining explanations from AI chatbots may amplify IOED while being associated with less effective retention of explanatory knowledge. Our findings are consistent with previous evidence that using Internet resources leads people to overestimate their explanatory knowledge (Fisher et al., 2015; Eliseev & Marsh, 2023) and negatively affects metacognitive predictions of task performance (Fisher et al., 2022). At the same time, they extend previously documented effects to situations of seeking explanatory information through chatbots.

Further research should focus more deeply on the mechanisms underlying this effect. We identify two possible components, each of which could potentially contribute to the IOED, or operate jointly. First, there is the effect of information misattribution, which in this context can be reformulated as the misattribution of explanatory knowledge. Fisher and Oppenheimer (2021) showed that when solving quizzes collaboratively with human and non-human assistants, the accuracy of predictions increases if participants are explicitly shown which number of correct answers was given by them and which by the assistants. Consequently, if misattribution of explanatory knowledge occurs during interaction with AI, the experimental design should include collaboration with AI for co-producing explanations, followed by indicating which parts of the explanation were improved by the AI. Second, the effect may be related to the "typing animation," i.e., the format of text presentation characteristic of modern chatbots, as presentation format is known to significantly influence reading performance (Benedetto et al., 2015). Such a format may also make the AI appear more anthropomorphic (Zhou et al., 2025), increasing trust and enhancing the user's susceptibility to persuasion (Williams-Ceci et al., 2024). If it creates the impression of communicating with a living person, it may further trigger effects associated with overreliance on external sources of knowledge within a community (see, for example, Sloman & Rabb, 2016).

Moreover, the question of individual differences that may moderate and potentially help mitigate this effect remains important. For example, metacognitive sensitivity has been shown to be an important predictor of confidence ratings in intelligence testing (Choo & Double, 2025), while cognitive reflection helps reduce IOED in socially desirable domains (Gaviria & Corredor, 2021), assess the pragmatic functions of explanations (Aslanov et al., 2025), and promote conceptual understanding of scientific information (Young & Shtulman, 2020). Therefore, the role of variables related to broadly understood rationality and metacognitive awareness should also be further examined in this context.

We also wish to acknowledge several limitations of the present study. First, given the relatively small effect sizes and the modest sample size, the observed results should be replicated with a larger number of participants. Second, although the texts in the GPT and no-GPT conditions were identical, participants in the GPT condition were required to remember and follow more complex instructions, holding them in working memory (i.e., entering questions into the chat rather than simply reading pre-generated texts). It remains unclear to what extent this additional cognitive demand may have affected text comprehension and subsequent explanation generation. However, future research could employ a within-subjects design to balance potential instruction-related effects. In addition, in the present study, the explanations primarily required mechanistic knowledge, whereas the effect may depend on the type of explanation. Therefore, including other types of explanations, such as teleological explanations, could help clarify the generalizability of this effect. Finally, it may be interesting to compare the GPT condition with a condition in which participants interact with human experts, in order to examine whether the observed effects are driven by exposure to structured and well-formulated explanations provided by "agents" more generally, rather than by AI systems per se.

Data availability statement

The materials as well as the data and scripts are available online: https://osf.io/kgpcx/?view_only=fc02a0e624d34d4791bd0f1a526d9357.

Acknowledgments

Funding from ANID Support 2024 AFB240004 is gratefully acknowledged. Partially supported by FONDEF ID24I10105 project. Funding from CMM Grant/Award FB210005 and CIAE Grant/Award FB0003 andAFB240004, ANID/PIA Basal Funds for Centers of Excellence are gratefully acknowledged.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: a construal level account of the illusion of explanatory depth. Journal of personality and social psychology, 99(3), 436.

Aslanov, I., Kotov, A., Guerra, E., Fedoriaieva, A., & Kotova, T. (2025). Seeking the category: The pragmatic function of formal explanations and the role of cognitive reflection. Cognitive Science, 49(8), e70101.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68(3),* 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1),* 1-48. https://doi.org/10.18637/jss.v067.i01

Benedetto, S., Drai-Zerbib, V., Pedrotti, M., Tissier, G., & Baccino, T. (2015). The effect of different text presentation formats on eye movement metrics in reading. Computers in *Human Behavior, 51*, 1184–1192. https://doi.org/10.1016/j.chb.2015.05.043

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3(30),* 774. https://doi.org/10.21105/joss.00774

Carroll, J. B. (1964). Language and thought. Prentice-Hall.

Choo, R., & Double, K. S. (2025). Metacognitive sensitivity moderates reactivity to confidence ratings. *Thinking & Reasoning.* https://doi.org/10.1080/13546783.2025.2542252

Christensen R (2023). *_ordinal-Regression Models for Ordinal Data_*. R package version 2023.12-4.1, <https://CRAN.R-project.org/package=ordinal>.

Eliseev, E. D., & Marsh, E. J. (2023). Understanding why searching the internet inflates confidence in explanatory ability. *Applied Cognitive Psychology, 37(4),* 711-720.

Fisher, M., & Keil, F. C. (2016). The curse of expertise: When more knowledge leads to miscalibrated explanatory insight. *Cognitive science, 40(5),* 1251-1269.

Fisher, M., & Oppenheimer, D. M. (2021). Who knows what? Knowledge misattribution in the division of cognitive labor. *Journal of Experimental Psychology: Applied*, 27(2), 292.

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, 144(3), 674.

Fisher, M., Smiley, A. H., & Grillo, T. L. (2022). Information without knowledge: The effects of Internet search on learning. In *Memory Online* (pp. 7-19). Routledge.

Gaviria, C., & Corredor, J. (2021). Illusion of explanatory depth and social desirability of historical knowledge. *Metacognition and Learning, 16(3),* 801-832.

Gaviria, C., & Corredor, J. (2025). Understanding, fast and shallow: Individual differences in memory performance associated with cognitive load predict the illusion of explanatory depth. *Memory & Cognition, 53(3),* 881-895.

Greussing, E., Guenther, L., Baram-Tsabari, A., Dabran-Zivan, S., Jonas, E., Klein-Avraham, I., ... & Song, H. (2025). The perception and use of generative AI for science-related information search: Insights from a cross-national study. *Public Understanding of Science, 34(5),* 599-615.

Guerra, E., Peña, M., & Araya, R. (2025). The Inevitable and Unpredictable Role of Large Language Models in Education: A Commentary on Huettig and Christiansen (2024). Cognitive Science, 49(8), e70105.

Huettig, F., & Christiansen, M. H. (2024). Can large language models counter the recent decline in literacy levels? An important role for cognitive science. *Cognitive Science, 48(8),* e13487.

Körner, R., Schütz, A., & Petersen, L. E. (2024). "It doesn't matter if you are in charge of the trees, you always miss the trees for the forest": Power and the illusion of explanatory depth. *Plos one, 19(4),* e0297850.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82(13),* 1-26. https://doi.org/10.18637/jss.v082.i13

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627(8002),* 49-58.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology, 87(1),* 1-32.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rinker, T. W. (2018). *textclean: Text cleaning tools (Version 0.9.3)* [Computer software]. https://github.com/trinker/textclean

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Selivanov, D., Bickel, M., & Wang, Q. (2020). *text2vec: Modern text mining framework for R [Computer software]*. https://CRAN.R-project.org/package=text2vec

Sloman, S. A., & Rabb, N. (2016). Your Understanding Is My Understanding: Evidence for a Community of Knowledge. *Psychological Science, 27(11), 1451-1460*. https://doi.org/10.1177/0956797616662271

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141-188. https://doi.org/10.1613/jair.2934

Ward, A. F. (2021). People mistake the internet's knowledge for their own. *Proceedings of the National Academy of Sciences, 118(43)*, 1–10. https://doi.org/10.1073/pnas.2105061118

Williams-Ceci, S., Zalmanson, L., MACY, M., & Naaman, M. (2024, September 14). Stereo-Typing: LLM Chatbots' Appearance of Typing can Increase Belief in a False Stereotype. https://doi.org/10.31234/osf.io/eqjsg

Young, A. G., & Shtulman, A. (2020). Children's cognitive reflection predicts conceptual understanding in science and mathematics. *Psychological Science, 31(11),* 1396-1408.

Zhou, D., Gallagher, J. R., & Sterman, S. (2025). Thoughtful, confused, or untrustworthy: How text presentation influences perceptions of AI writing tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery. https://doi.org/10.1145/3698061.3726907