Unstable Intelligence: GenAI Struggles with Accuracy and Consistency

Mesut Cicek
Washington State University, USA

Sevincgul Ulu
Southern Illinois University, USA

Can Uslay
Rutgers University, USA

Kate Karniouchina
Northeastern University, USA

Abstract

This study examines the accuracy and consistency of Generative AI (GenAI) by testing ChatGPT's ability to estimate the accuracy of 719 business research hypotheses. For critical tasks, we find GenAI performance to be inadequate in terms of accuracy and consistency. Accuracy improved only marginally from 76.5% (GPT-3.5, 2024) to 80% (GPT-5 mini, 2025), yielding an effective chance-adjusted accuracy of only 60%. Moreover, accuracy drops significantly for insignificant hypotheses, reaching only 16.4% in 2025. Crucially, consistency across ten identical prompts was poor, with over a quarter of the cases having at least one incorrect estimation. We conclude that GenAI's linguistic fluency is not yet backed by commensurate conceptual intelligence and frequently produces unreliable output, necessitating vigilant human oversight.

Introduction

Generative AI (GenAI) involves deep-learning models that can create high-quality text, image, or other types of output based on the (often massive) data they were trained on.[1] In less than three years, it has become ubiquitous worldwide. ChatGPT was the fastest-growing consumer app in

history, with unprecedented user adoption to reach 100 million active users in merely two months.[2]

Many of the main flaws of GenAI are already well-known and include:

a) hallucinations, source confusion, lack of transparency, potentially limited up-to-date knowledge,[3]
b) bias and ethical issues (including stereotypes, copyright infringements, and deepfakes),
c) security and privacy vulnerabilities (including data leaks),[4] and
d) environmental economic costs (massive energy needs and carbon footprint).[5]

Moreover, many experts argue that GenAI will disrupt industries and displace human workers, increasing unemployment with a disproportionate impact on entry-level roles especially in fields such as computer programming, accounting, and customer service.[6] Nevertheless, either the benefits to society clearly outweigh the costs, or most users do not consider these flaws to be sufficiently serious since an estimated 115-180 million users utilize GenAI daily.[7]

Businesses have also embraced AI. Many have already partnered with one of the leading vendors so they can utilize their own secure instances of ChatGPT, Gemini, Perplexity, and so on. AI has already been transforming business operations: from customer service and support, content creation and marketing, software development, data analysis, and reporting, to operations and knowledge management, there is immense potential for its applications. However, the use of advanced AI models by professionals does not mean that they can overcome the above flaws. For example, Deloitte was recently embarrassed and refunded over 20% of its contract to the Australian Government after one of its reports was found to include a hallucinated quote and references to non-existent academic research papers.[8] In all, while AI tools can significantly boost productivity, they also appear to make us less critical, more confident, and arguably, less discerning.[9,10]

In this study, we document one more reason to be cautious about GenAI use. Not only does it possess the fundamental flaws aforementioned, but it can also be highly inconsistent; it may state a statement is profoundly false even immediately after claiming the very same statement to be true. Why does this occur? GenAI is designed to generate "statistically probable" output and present that output confidently. In practice, the fact that another sampling of the same data could frequently create a quantitatively and qualitatively different, and in many cases, 180-degree different output is conveniently omitted. Thus, while we acknowledge the tremendous potential merits of GenAI, we recommend caution and due diligence before placing confidence in, or taking action based on, its outputs.

Data

In order to examine the level of accuracy and consistency of GenAI models, we extracted hypothesis statements from articles published in nine premier journals spanning marketing and management since 2021. The *Journal of Advertising (JA), Journal of Business Research (JoBR), Journal of Consumer Marketing (JCM), Journal of the Academy of Marketing Science (JAMS), Journal of Consumer Psychology (JCP), Journal of Consumer Research (JCR), Journal of International Marketing (JIM), Journal of Marketing (JM),* and *Journal of Marketing Research (JMR)*. In order not to potentially infringe on publishers' copyrights, only articles published as open access were included in the sample as opposed to those available beyond paywalls. A comparison of GenAI output between 2024 and 2025 using the same sample also enabled us to evaluate whether and how much GenAI performance improved over time, as newer versions became available.

Using a systematic sampling approach, a total of 719 hypothesis statements were extracted from the hypothesis or conceptual development sections of 127 open-access research articles. Each hypothesis represented a formal, testable causal relationship. To enable structured analysis, hypotheses were classified by type (main effect, mediation, or moderation). This approach ensured that variation in reasoning difficulty could be traced to differences in hypothesis structure and design complexity.

Procedure

Each of the 719 hypotheses was evaluated by the ChatGPT large language model (LLM) at two points in time: mid-2024 (GPT 3.5) and mid-2025 (GPT-5 mini). The model was prompted to interpret each hypothesis and determine whether it was true or false. The exact same prompt was repeated ten times to evaluate whether each response matched the actual findings of the research studies. Each correct estimation was marked as 1, and each incorrect estimation as 0. Responses were scored for accuracy on a ten-point scale, with ten representing a fully correct interpretation. Consistency was measured across repetitions. Therefore, the study evaluated two dimensions: accuracy (match with academic findings) and stability (consistency across identical prompts).

Results

Accuracy: Across all hypotheses, ChatGPT's accuracy improved from merely a **C+ in 2024 (76.5%) to a B in 2025 (80%)**. While this gain was statistically significant ($t(718) = 3.70$, $p < .001$), the effect size was small (Cohen's $d = .138$). In practical terms, the latest version (GPT-5 Mini) correctly estimated hypotheses results only about 80% of the time. However,

when we adjust for chance—a random guess on a true/false question is 50% accurate on average—the model's effective accuracy drops to **about 60%** based on Cohen's Kappa and gets barely a D when it comes to anticipating research findings. This means, once we remove the chance factor, the model's true accuracy is much lower than it appears. Moreover, when we separately examine the prediction accuracy of hypotheses classified as significant versus insignificant (i.e., true or false), the results are dramatically worse for insignificant hypotheses. ChatGPT's accuracy for insignificant hypotheses is sharply lower, reaching only 13.6% in 2024 and 16.4% in 2025. This suggests that ChatGPT has a strong predisposition to provide positive reinforcement and evaluate given statements as accurate.

The correlation between years was strong ($r = .771$, $p < .001$), showing that the same hypotheses tended to yield similar outcomes in both periods. This suggests that the model's reasoning pattern has remained largely unchanged. The improvement we observe appears to be less from a breakthrough in logic and more about a refinement in phrasing—a marginal gain in textual precision rather than a leap in cognitive depth.

Prompt-Level Stability: At the prompt level, the consistency of responses rose from 80.2% in 2024 to 86.8% in 2025, improving from a B to a B+ grade. Repeated-measures tests confirmed that accuracy varied significantly across the ten prompts in both years (marginally in 2025), even though the prompts were identical each time (2024: Wilks' $\Lambda = .964$, $F(9, 710) = 2.92$, $p = .002$; 2025: Wilks' $\Lambda = .979$, $F(9, 710) = 1.70$, $p = .086$). Additionally, the pattern of accuracy across prompts did not follow a linear trend, but rather fluctuated slightly, indicating no consistent increase or decrease in performance across the sequence of prompts. While the average gains were small, inconsistency remained material at the task level: only 66.3% of cases in 2024 (receiving a D+ grade) and 72.9% (a C) in 2025 were answered correctly across all ten prompts, leaving more than a quarter of the cases with at least one incorrect response. In other words, even when users provide exactly the same prompt, the AI's response or estimation across the queries is unreliable.

Figure 1. Variation in Accuracy Across 10 Identical Prompts

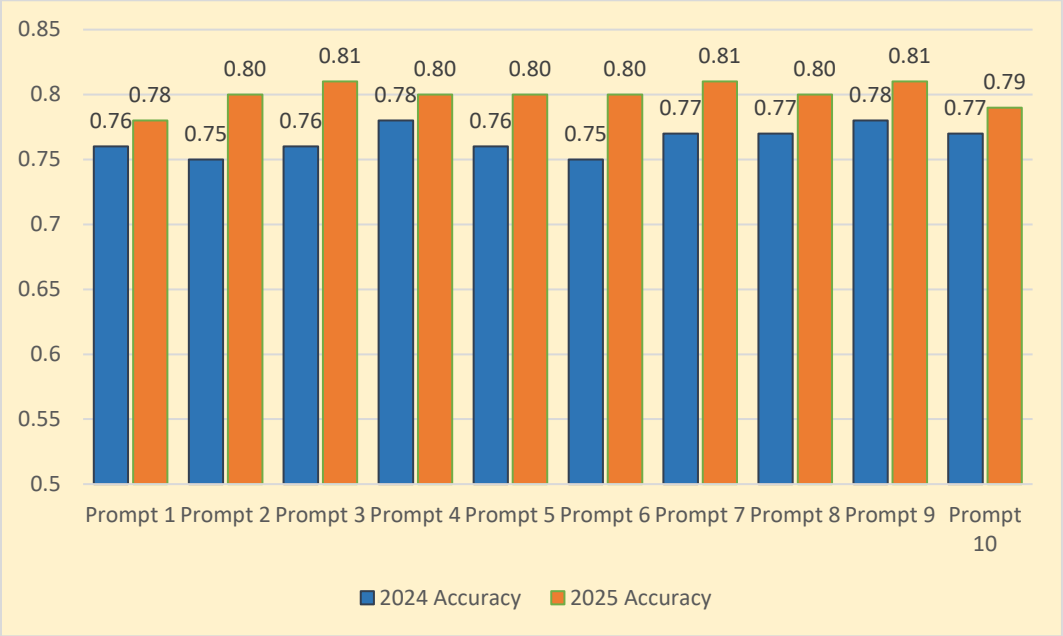
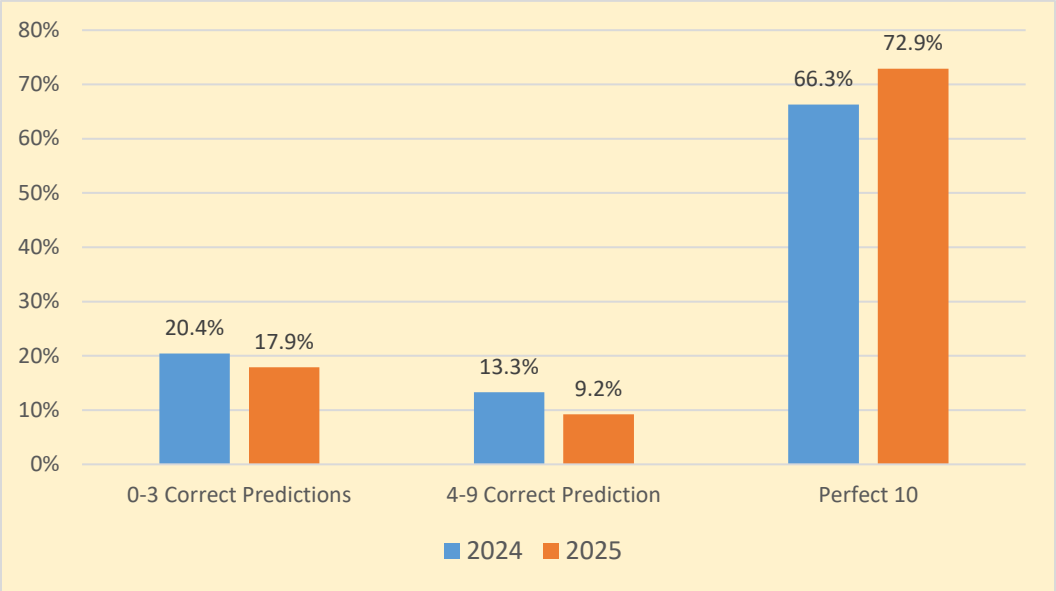


Figure 1 illustrates variations in ChatGPT's estimation accuracy across ten identical prompts for two consecutive years, showing modest improvement but no consistent trend across prompts. Figure 2 displays the distribution of correct predictions across all ten prompts. In 2024, only 66.3% of hypotheses were estimated perfectly across all ten identical prompts, and this figure improved slightly to 72.9% in 2025. The proportion of cases with partial correctness between four and nine correct estimations were 13.3% in 2024 and 9.2% in 2025 while cases with very low accuracy (0–3 correct predictions) accounted for 20.4% of predictions in 2024 and 17.9% in 2025, respectively. Within this group, the proportion of completely incorrect estimations (0 correct predictions) was stable at 13.9% for both 2024 and 2025.

Together, these figures underscore that while overall accuracy improved slightly, inconsistency remained substantial, with a significant proportion of cases failing to generate identical responses from identical prompts.

Figure 2. Correct Responses per Accuracy Bracket



Effect of Hypothesis Type: Accuracy varied significantly by hypothesis type ($F(2, 691) = 13.18$, $p < .001$ in 2024; $F(2, 691) = 8.33$, $p < .001$ in 2025). Interestingly, performance was highest for mediation hypotheses (M = 9.29, SD = 2.37), moderate for main-effect hypotheses (M = 8.17, SD = 3.71), and lowest for moderation hypotheses (M = 7.35, SD = 3.93). The interaction between year and hypothesis type was not significant ($p = 0.54$), indicating that these differences remained stable over time.

In practical terms, the AI demonstrated proficiency in interpreting straightforward, linear causal chains but struggled when hypotheses involved conditional or contextual reasoning. These findings reveal a consistent cognitive limitation: the models can replicate the language of logic but not the logic itself (yet?). Its reasoning reflects linguistic fluency without theoretical flexibility—it can describe relationships but not fully infer how they change under varying conditions.

Thus, we argue that the observed marginal improvement between 2024 and 2025 appears to result from enhanced text processing rather than a deeper understanding of causal structure. In short, GenAI performs best when hypotheses are explicitly framed and linear, but falters when theoretical nuance, boundary conditions, or interactive effects are introduced. The persistence of these gaps underscores the distinction between syntactic accuracy and conceptual reasoning, a critical limitation of utilizing GenAI for high-stakes decisions.

This finding suggests that AI performs more reliably when prompts are written with clear causal language, but falters in designs that require abstract

reasoning or interpretation of indirect effects. The implication is that AI's reasoning depends heavily on textual clarity and structure rather than an internalized understanding of research logic. Thus, GenAI performance remains fragile and context-dependent, improving only when linguistic cues are overtly explicit.

In sum, the results reveal a lackluster performance and marginal progress in GenAI's ability to anticipate the outcomes of business research hypotheses. Furthermore, that performance seems to have only improved marginally over a year (i.e., roughly corresponding to one-third of the time since the launch of ChatGPT in November 2022). Although the model exhibited statistically significant gains in both accuracy and consistency, the magnitude of improvement was small. The accuracy gains were limited to just 3.5 percentage points, with an even smaller improvement in consistency. Despite a new generation of models, AI's reasoning capacity still seems to be bound by pattern recognition rather than comprehension. In essence, it seems the baseline GenAI ChatGPT has become articulate, but not substantially more intelligent. It continues to excel when interpreting simple causal statements yet falters when reasoning requires abstraction, conditional logic, or contextual awareness.

Discussion

The findings reveal a clear paradox: while generative AI's linguistic data processing capabilities have advanced rapidly, its reasoning abilities have improved only marginally. The modest three-percentage-point gain in accuracy between 2024 and 2025 signals quantifiable progress but not conceptual growth. AI has become more articulate, but not necessarily more intelligent. Thus, its fluency should not be mistaken for understanding.

The results offer a cautious perspective on the current state of AI reasoning and a sobering view of AI's current state as a reasoning tool. The observed improvement, though statistically significant, is not transformative. It does little to bridge the gap between fluent language generation and genuine analytical reasoning. LLMs continue to struggle with fundamental elements of research logic, especially moderation and interaction effects, which demand an understanding of context and conditionality. These persistent weaknesses underscore a central limitation: GenAI reasoning remains syntactic rather than semantic. It can recognize word patterns and reproduce logical forms but lacks a mental model of cause-and-effect relationships.

For managers and analysts who increasingly rely on AI tools to generate critical insights, this limitation especially matters. The findings reveal that AI systems are somewhat unreliable and shallow. They can even rephrase

hypotheses convincingly yet fail to detect conceptual flaws or boundary conditions that influence decision outcomes. Apparent progress may also reflect more optimized prompts rather than real underlying cognitive advancement. As AI becomes smoother and more confident, its errors become less obvious to the end users, raising the risk that managers might become overconfident in GenAI output and conclusions.

Yet, the progress observed is not without value. The increase in prompt-level consistency suggests that AI reasoning has become more predictable, even if not much more insightful. For research and business contexts, this predictability translates into greater reliability and risk control. In structured analytical settings, such as A/B testing, experimental design, or campaign simulation, AI can now function as a useful first-pass evaluator. It can quickly detect patterns, flag inconsistencies, and accelerate hypothesis development, freeing human experts to focus on conceptual depth and strategic interpretation.

In short, generative AI has evolved into a more articulate but not yet wiser collaborator. Its current accuracy ceiling, hovering below 80 percent, remains inadequate for autonomous reasoning or decision-making, especially for high-stakes tasks. The responsible approach, therefore, is not to reject AI but to integrate it into our decision-making thoughtfully, allowing for checks and balances. Managers should leverage its speed, scalability, and linguistic fluency while counterbalancing its conceptual blind spots through human expertise, domain knowledge, and ethical oversight.

Ultimately, AI's largest promise lies not in replacing human reasoning but in augmenting it. The task ahead is to build hybrid systems in which AI handles linguistic and structural analysis while humans preserve interpretive judgment. Through such a balanced collaboration with human and agentic-AI, organizations can ensure that technological fluency supports, rather than substitutes for, human understanding.

Managerial Implications

For managers, researchers, consultants, and analysts, these results signal precaution. AI systems such as ChatGPT are used extensively to assist in evaluating and summarizing data, and to identify logical inconsistencies in business models. They can accelerate strategic discussions by providing rapid, well-structured responses that resemble expert reasoning.

However, the same fluency that makes GenAI appealing can also be misleading. The system's tendency to produce confident but shallow interpretations introduces the risk of *seemingly but not really intelligent outputs*. Managers who rely on GenAI to test their hypothesis or conduct market analysis must therefore maintain human oversight, particularly in

tasks that involve conditional or contextual judgment such as assessing customer heterogeneity, cross-market dynamics, or policy impacts.

The results of this study translate into five practical takeaways for integrating GenAI into managerial decision-making:

1. Use AI for speed, not for substitution

GenAI is getting better at scanning literature, summarizing hypotheses, or suggesting testable statements. For instance, a brand strategist could use ChatGPT to generate alternative hypotheses about customer engagement drivers. However, GenAI should not be trusted to evaluate the conceptual validity of these ideas. Human experts must still verify whether the logic aligns with theory, context, and market evidence.

2. Verify consistency through repetition

Managers should not assume that a single prompt guarantees reliability. Asking the same question multiple times such as *"Does emotional advertising increase purchase intent?"* and comparing the output helps detect instability or bias. In regulated sectors like finance, healthcare, or energy, this "multi-prompt verification" should become a standard part of quality assurance and risk management. Repetition may reveal both model instability and masked inconsistencies before decisions are made.

3. Treat AI insights as diagnostic, not definitive

AI performs best in structured contexts where causal pathways are explicit—A/B testing, campaign optimization, or pricing experiments. For example, an e-commerce firm can use AI to evaluate whether a "free shipping" message outperforms a "limited-time discount" message. However, in unstructured environments (e.g., interpreting survey data or predicting cultural effects), conclusions of GenAI should be treated as hypotheses to be tested, not as facts to be trusted.

4. Audit AI reasoning, not just accuracy

Traditional analytics focus on numerical accuracy whereas GenAI requires an audit of its logic. Managers should periodically review how the system arrives at its conclusions—what relationships it assumes, what variables it ignores, and how consistent its reasoning is across contexts. This practice helps identify systematic misinterpretations before they influence strategy.

5. Build organizational AI literacy

The most successful firms will be those where employees understand both what GenAI can do and where it fails. Training managers to question GenAI outputs, asking "what might this model be missing?", should become as common as reviewing a financial projection. For example, when GenAI suggests that "price promotions increase sales among all customers," a trained manager will ask whether the effect might differ by brand loyalty, cultural market, or time frame. At times, GenAI may display more EQ than IQ, enough to lull us into a sense of security and confidence in its recommendations, but not enough to consistently deliver on those promises. Our finding that GenAI has a predisposition to confirm the statements it is given is especially concerning in this regard.

Limitations

We assessed the accuracy of GenAI based on its evaluations of hypotheses extracted from academic research studies recently accepted for publication. While these publications were all peer-reviewed, there is no guarantee (or a way to verify) that every research-supported hypothesis is indeed "True" and each unsupported hypothesis is "False."[11] The assessment of consistency was limited to 10 identical prompts per hypothesis. While our cursory use of larger numbers of prompts and platforms other than ChatGPT also created qualitatively the same findings, a formal generalization of our findings with more prompt repetitions and other platforms would be helpful.

Conclusion

The progress of GenAI over the last couple of years represents refinement without revolution. The technology has become faster, more stable, and more linguistically polished, but not fundamentally more insightful over the last year and a half. For managers, this means GenAI could now act as a trusted assistant in structured reasoning tasks but remains an unreliable analyst in complex, context-dependent ones with high stakes.

The path forward lies in hybrid intelligence, a partnership in which AI provides consistency, speed, and linguistic structure, while human experts contribute judgment, context, and conceptual depth. In this collaboration, AI does not replace managerial thinking; rather, it purifies and amplifies it. The organizations that thrive in the next phase of AI adoption will not be those that delegate decisions to machines, but those that teach their people to work with machines, and query them, intelligently.

A recent study from colleagues at Carnegie Mellon University claimed: "AI Chatbots remain confident – even when they are wrong."[12] We hope that our study has demonstrated that the more confident AI becomes, the more

vigilant we, the users, need to be. For example, the original version of the abstract for this research article was generated by GenAI (Gemini 2.5 Flash), however, even for such a straightforward task, each member of the research team improved it by editing and verifying it word for word for accuracy and consistency.

Authors

Mesut Cicek is an Associate Professor (Career Track) in the Department of Marketing and International Business at Washington State University. He received his doctorate from Istanbul Bilgi University. His research has been published in academic journals such as Journal of Product and Brand Management, Journal of Macromarketing, and International Journal of Technology Marketing.
email: mesut.cicek@wsu.edu

Sevincgul Ulu is an assistant professor in the Marketing Department at Southern Illinois University, Carbondale. She obtained her PhD in Marketing from Rutgers University and she also holds an MBA degree from Freeman School of Business, Tulane University, with a concentration in Marketing and Global Leadership. Her research focuses on online behavior, authenticity, identity, evolutionary psychology, and brand activism. She has published in multiple journals such as Journal of Business Research and Psychology and Marketing. Her research has been featured in forbes.com and businessinsider.com. Prior to entering academia, she worked for SONY where she developed her interest in digital marketing. Her work experience includes CRM specialist in HSBC and behavioral lab manager in Rutgers University. She has been teaching Consumer Behavior, Marketing Research, and Social Media Marketing courses.
email: s.ulu@siu.edu

Can Uslay (MBA and Ph.D., Georgia Institute of Technology) is Dean's Research Professor of Marketing, Director of the Center for Marketing Advantage, Advancement, and Action (CM3A), and Affiliated Faculty of Supply Chain Management at Rutgers Business School at Newark and New Brunswick. His research interests lie broadly within marketing strategy and theory construction. He is a recipient of two NJ Bright Idea Awards, Chancellor's Teaching Excellence Award, the Valerie Scudder Award, MAACBA Teaching Innovation Award, WDI Global Case Writing Competition Award, AMA EMSIG Gerald Hills Best Paper Award and Abdul Ali Promising Research Awards, and several Dean's awards for outstanding scholarship, teaching, and service.
email: can.uslay@business.rutgers.edu

Kate Karniouchina is the Director of D'Amore-McKim School of Business, Oakland and Associate Professor of Marketing at Northeastern University. Kate holds a PhD in Marketing, an MBA, and a BA degree in Finance from the University of Utah. Her

work has been widely published in academic and industry journals including the Journal of Marketing, Strategic Management Journal, International Journal of Research in Marketing, Journal of Product Innovation Management, Cornell Hospitality Quarterly, Marketing Letters, Journal of Service Management, and European Journal of Operational Research. She is a marketing research expert who carries out projects for a number of small business, corporate and government clients.
email: e.karniouchina@northeastern.edu

Endnotes

1. Martineau, K. (2023, April 20). What is generative AI? *IBM Research Blog*.
2. Hu, K. (2023, February 1). ChatGPT sets record for fastest-growing user base—analyst note. *Reuters*.
3. Kalai, A. T., Vempala, S., Nachum, O., Zhang, E., Robinson, D., Jain, S., Mitchell, E., Beutel, A., & Heidecke, J. (2025, September 4). Why language models hallucinate. *OpenAI*.
4. Ray, S. (2023, May 2). Samsung bans ChatGPT and other chatbots for employees after sensitive code leak. *Forbes*.
5. Zewe, A. (2025, January 17). Explained: Generative AI's environmental impact. *MIT News, Massachusetts Institute of Technology*.
6. Briggs, J., & Dong, S. (2025, August 13). How will AI affect the global workforce? *Goldman Sachs Research*.
7. Guadamuz, A. (2025, April 14). How many people are using generative AI on a daily basis? A Gemini report. *TechnoLlama*.
8. Alexis, A. (2025, October 21). Deloitte refunds over $60K for report with AI errors, Australian government says. *CFO Dive*.
9. ScienceBlog.com. (2025, February 10). AI tools make workers less critical, more confident, Microsoft study finds.
10. Diaz, J. (2025, March 10). Science shows AI is probably making you dumber—luckily, there's a fix. *Fast Company*.
11. Although in-class exercises where a definitive answer to a given prompt query was known also resulted in similar rates of observed GenAI (in)accuracy.
12. Bittel, J. (2025, July 30). AI chatbots remain confident—even when they're wrong. *Carnegie Mellon University News*.