Generative Chatbots AIn't Experts: Exploring Cognitive and Metacognitive Limitations that Hinder Expertise in Generative Chatbots.

Trent N. Cash* and Daniel M. Oppenheimer

Department of Social and Decision Sciences, Carnegie Mellon University, USA

Department of Psychology, Carnegie Mellon University, USA

Author Note

Correspondence concerning this article should be addressed to Trent N. Cash, Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Ave., PH 222 Pittsburgh, PA, 15213. E-Mail: Tcash@andrew.cmu.edu, Phone: 330-451-9972, ORCID: 0000-0001-6624-9616. Because this was a commentary, no data, materials, or analysis code were used. The authors declare no financial support nor any conflicts of interest.

Abstract

Despite their ability to answer complex questions, it is unclear whether generative chatbots should be considered experts in any domain. There are several important cognitive and metacognitive differences that separate human experts from generative chatbots. First, human experts' domain knowledge is deep, efficiently structured, adaptive, and intuitive – whereas generative chatbots' knowledge is shallow and inflexible, leading to errors that human experts would rarely make. Second, generative chatbots lack access to critical metacognitive capacities that allow humans to detect errors in their own thinking and communicate this information to others. Though generative chatbots may surpass human experts in the future – for now, the nature of their knowledge structures and metacognition prevent them from reaching true expertise.

Keywords: Generative Chatbots; Artificial Intelligence; LLM; Expertise; Cognition; Metacognition;

Generative Chatbots AIn't Experts: Exploring Cognitive and Metacognitive Limitations that Hinder Expertise in Generative Chatbots.

Generative AI chatbots built on Large Language Models – such as ChatGPT, Gemini, and Claude – burst onto the public scene in 2022. Since then, one key area of discussion is the extent to which generative chatbots think and learn like humans.

Imundo et al. (2024) approach this question with a thoughtful review of the literature on the relationship between generative chatbots and expertise. They compare human expertise to artificial expertise, assess the benefits of generative chatbot usage for experts and non-experts, and evaluate the extent to which generative chatbots may enhance or inhibit the development of expertise among human users. One key distinction that was largely overlooked was the difference between the (meta-)cognitive capacities (what we know about what we know) of generative chatbots and human experts. We aim to fill this gap. First, we will discuss differences in how generative chatbots and humans develop and structure knowledge – and the implications for intuition and adaptability. Second, we will compare the metacognitive strategies that human experts and generative chatbots use to detect errors in their cognition and communicate the likelihood of errors to others. We argue that these differences are sufficiently stark that generative chatbots cannot be considered experts, at least not in their current forms.

Developing Expertise

For humans, developing expertise involves mastering a series of steps, each of which is necessary to master the subsequent step (Ericsson et al., 1993; Van De Pol et al., 2010). Consider a student learning to play piano. First, they learn how to play notes. Next, they learn how to play

chords. Then they learn how to read sheet music. Then they apply these skills to increasingly complicated compositions, until one day they are playing Beethoven. Through deliberate practice (Ericsson et al., 1993), the student establishes a base of expertise that is built on a foundation of lower-level skills. For this reason, it would be surprising to meet a pianist who can play a sonata but can't play a simple chord.

This pattern can also be applied to declarative knowledge. When humans learn concepts, we integrate them into schemata – associative mental frameworks that connect related ideas (Derry, 1996; Rawson & Van Overschelde, 2008). These schemata are often hierarchical, built on a base of lower-level knowledge that can be used to better explain and understand more complex, higher-level concepts (Ghosh & Gilboa, 2014; Rumelhart & Ortony, 1977). To develop effective schemata, however, the lower-level knowledge must be acquired first. It would be quite difficult to understand how diabetes works without knowing what insulin is.

In contrast, generative chatbots do not develop expertise in a sequential fashion. They are built on large language models that simply form associative neural networks that reflect how often certain words or phrases co-occur in their training set (Wu et al., 2023). As such, generative chatbots may develop associative schema of how concepts are related (e.g., monkeys and bananas frequently co-occur) but are often missing foundational links within their schemata. This results in strange outcomes where generative chatbots can exceed at complex tasks but fail at simple ones. For example, ChatGPT can solve undergraduate math problems (Frieder et al., 2023) but frequently fails when attempting to do simple addition (Cheng & Yu, 2023).

To the extent that we consider deep, structural knowledge to be a criterion of expertise, this kind of error demonstrates that generative chatbots are not experts – but rather simulations that mimic what an expert might say in response to a question. True experts would rarely, if ever, fail at the

most basic tasks within their field. As such, we believe that assessing the expertise of generative chatbots is best approached through a strategy presented by Shlomi Sher (2023): “Don’t focus on the most complex things the system can do. Instead, focus on the simplest things it can’t do.”

Expertise and Intuition

A second order effect of generative chatbots’ lack of deep structural knowledge is that – unlike human experts (Salas et al., 2010) – generative chatbots are unlikely to develop intuition about a domain. When humans develop expertise, cognitive processes that were formerly effortful become automatic (Salas et al., 2010; Samuels & Flor, 1997) and connections between concepts in schemata are reinforced, increasing the speed and efficiency with which relevant information can be recalled (Grabner et al., 2006; Huet & Mariné, 2005).

Using this developed intuition, human experts often make accurate inferences on far fewer data points than a generative chatbot requires (Salas et al., 2010; Wu et al., 2023). This allows humans to operate relatively efficiently in low-information environments (Ahn et al., 1992; Shanteau, 1992) and apply existing knowledge structures to solve novel problems (Christopher & Müller, 2014). For example, imagine that an expert chess player and an AI tool were asked to play an altered version of chess where the King can move two spaces (instead of one, as in standard chess). The human expert would be able to adapt their existing knowledge to accommodate this rule change, while the generative chatbot would likely struggle because this version of chess does not exist in their training set.

This limitation of generative chatbots is well-highlighted by tests that are used to detect bots, including generative chatbots. For example, Rodriguez and Oppenheimer (2024) demonstrated that when bots are shown a picture of a snow sculpture depicting feet and asked what would

happen to the feet on a warm day, they respond as if the feet were human feet – often stating that they will sweat. This is likely because most feet in their training set belonged to humans, so the association they make is warmth + feet = sweat. In contrast, humans can use their adaptive knowledge of the world to recognize that the image does not depict human feet, but rather snow. As this demonstrates, bots lack the cognitive flexibility to interpret abnormal and unexpected inputs.

Generative chatbots will almost certainly overcome these specific tests in the near future. For this reason, computer scientists are actively working to develop tests that can more effectively detect generative chatbots (e.g., Deng et al., 2024). Regardless, the fact remains that bot-detection strategies highlight the difference between human expertise – which is flexible and adaptive – and AI expertise, which is relatively inflexible and dependent on its specific training set.

Error Monitoring

Of course, this is not to say that human intuition is flawless or that humans are infinitely adaptable. Experts, like all humans, are prone to overconfidence (McKenzie et al., 2008) and often make erroneous judgments that are worse than those that would be made by a simple algorithm (Dawes et al., 1989). However, one sign of expertise is the ability to detect when you have made an error in your cognitive processing – known in the metacognition literature as error monitoring (Patel et al., 2011; Yeung & Summerfield, 2012)

The first line of defense in error monitoring is passive. When engaging in any sort of cognition, humans subconsciously evaluate several experiential factors about their cognitive process – such as how easy it was to retrieve the necessary information (Thompson et al., 2013), how easily

they were able to perceive the relevant stimuli (Reber & Schwarz, 1999), or how familiar aspects of the task felt (Alter & Oppenheimer, 2008). These cues generate a sense of *fluency*, a subjective experience of how easy it was to engage in the task (Alter & Oppenheimer, 2009; Oppenheimer, 2008). When fluency is low, humans can often tell that something just doesn't *feel* right.

Of course, fluency is not always an accurate cue. For example, there is evidence that learners use fluency as a signal that that they have learned material well, even when fluency is not actually predictive of performance (Ackerman & Zalmanov, 2012; Finn & Tauber, 2015). However, there are many cases in which disfluency can be a meaningful signal that something has gone awry and that additional consideration is needed (Ackerman & Zalmanov, 2012; Thompson et al., 2011).

Fluency may be a particularly useful cue for experts who have automatized a given task – and therefore expect high levels of fluency (Jiang & Hong, 2014; Schwarz, 2004). When a task feels disfluent and therefore violates their expectations, the expert will recognize that something is amiss (Roese & Sherman, 2007). Similarly, experts are likely to notice when a typically familiar task feels unfamiliar – thus inducing disfluency (Whittlesea & Williams, 1998). In contrast, a novice will likely not notice these sources of disfluency because they anticipate the task feeling disfluent due to their lack of expertise (Jiang & Hong, 2014). Relatedly, novices may be particularly likely to misinterpret incidental fluency as a signal that they have more expertise than they actually do (Finn & Tauber, 2015). Therefore, fluency provides a mechanism by which experts may be better at error monitoring than novices.

To our knowledge, there is no empirical evidence that generative chatbots experience (dis)fluency. On the surface, it seems reasonable to conclude that because generative chatbots do

not have feelings – which they readily admit when asked – they would not be able to experience fluency or disfluency in the way that humans do[1]. This lack of metacognitive feedback is a key difference between human experts and generative chatbots. While humans may overlook smaller errors that generative chatbots are good at detecting (e.g., typos), metacognitive disfluency allows humans to sense when something is just not right with the information that they are processing (Alter & Oppenheimer, 2009; Oppenheimer, 2008). For example, imagine analyzing an experimental dataset in which every participant gave the same response to a question. A generative chatbot would be unlikely to notice that anything was amiss, whereas an expert human analyst would notice that something was fishy (c.f. Simonsohn et al., 2024). In this sense, disfluency acts as a sort of backstop for human experts that generative chatbots lack.

Generating Metacognitive Judgments

While metacognitive judgments are rarely perfectly aligned with objective measures of performance, there is often a substantial correlation. For example, experimental participants are more likely to accurately solve problems they judge as more solvable (Burton et al., 2023), more likely to recall items that they judge as having learned better (De Bruin et al., 2007; Nelson & Dunlosky, 1991) and more often correct when their confidence is higher (Han & Dunning, 2024).

In many cases, experts make more accurate metacognitive judgments than novices. For example, high-performing students more accurately predict their performance on tests than low-achieving students (Hacker et al., 2000), expert chess players generate more accurate judgments of learning about opponents' strategies than novices (De Bruin et al., 2007), and content experts in the

[1] It is *possible* that generative chatbots, like humans, could experience disfluency when it takes longer to process information (Ackerman & Zalmanov, 2012). Despite this, there is no reason to believe that this disfluency would be associated with worse performance, thus making it a meaningless cue.

domains of climate science, statistics, and investment display less overconfidence than non-experts, as well as greater metacognitive knowledge (Han & Dunning, 2024). Of course, experts are still subject to common human biases, such as overconfidence (Einhorn & Hogarth, 1978), and there are domains in which experts' metacognitive judgments are no more accurate (but rarely less accurate) than those of novices (Cash & Oppenheimer, 2024; Han & Dunning, 2024; Lichtenstein & Fischhoff, 1977). However, on average there is a positive relationship between expertise and accuracy of metacognitive judgments.

To date, there is limited empirical evidence regarding the capacity of generative chatbots to make meaningful metacognitive judgments. That said, extant theory predicts the extent to which generative chatbots can use the types of cues that inform humans' metacognitive judgments. Koriat's (1997) cue-utilization approach distinguishes between three types of metacognitive cues: *intrinsic* cues derived from characteristics of the task itself (e.g., difficulty); *extrinsic* cues about the environment in which the task is presented (e.g., time limits); and *mnemonic* cues that reflect the individual's subjective, metacognitive experience (e.g., fluency).

Out of the three types of metacognitive cues, generative chatbots are best equipped to evaluate *intrinsic* cues for two reasons. First, *intrinsic* cues do not require any sort of introspection or self-awareness, which generative chatbots currently lack (Long, 2023). Second, *intrinsic* cues largely depend on statistical regularities that exist in the world – such as longer anagrams being harder to solve (Kaplan & Carvellas, 1968). Generative chatbots can learn these statistical relationships through their training set or can be explicitly trained to recognize these patterns. They can then use this knowledge to generate more accurate metacognitive judgments. In fact, generative chatbots may even be better than humans at detecting these statistical patterns and interpreting them as metacognitive cues.

Extrinsic cues are a more interesting case. On one hand, generative chatbots are unlikely to be impacted by many of the environmental factors that serve as cues for humans' metacognitive judgments, such as temporal delays (Nelson & Dunlosky, 1991) and prior experience (De Bruin et al., 2007; Han & Dunning, 2024). On the other hand, it is possible that generative chatbots could be aware of constraints that the environment places on their performance, such as time or word limits. This is particularly relevant given that OpenAI recently released a new generative chatbot – called OpenAI o1 – that purportedly spends more time thinking before providing an answer, with the goal of improving response quality.

Finally, as described in the section on fluency, it is unlikely that generative chatbots experience the metacognitive feelings necessary to base their metacognitive judgments on *mnemonic* cues. Moreover, generative chatbots would likely lack sufficient self-awareness to engage in the level of self-reflection necessary to interpret these cues (Long, 2023). As such, *mnemonic* cues are unlikely to have any impact on generative chatbots' metacognitive judgments.

The fact that generative chatbots have access to only a subset of the metacognitive cues that humans have access to leads to certain theoretical predictions. For example, generative chatbots should have relatively better metacognition for predictions involving aleatory uncertainty (i.e., uncertainty derived from inherent stochasticity) than predictions involving epistemic uncertainty (i.e., uncertainty derived by lack of knowledge, including knowledge about oneself; Fox & Ülkümen, 2011). This is because aleatory uncertainty is often reduced via *instrinsic* and *extrinsic* cues, whereas epistemic uncertainty is often reduced via *mnemonic* cues.

Indeed, Cash et al. (2024) recently conducted a study in which they asked humans and generative chatbots to make predictions about future events (e.g., football games and awards shows) then judge their confidence in those predictions. Because these events were in the future, uncertainty

about the outcomes of the events was inherently aleatory (Fox & Ülkümen, 2011). Generative chatbots' confidence judgments were as accurate as – and in some cases, more accurate than – those of humans. This aligns with predictions from the cue utilization model (Koriat, 1997), as both human participants and the generative chatbots were likely able to base their metacognitive judgments on *instrinsic* cues, such as how similar the records of the two football teams were.

However, Cash et al. (2024) also demonstrated that generative chatbots, especially Gemini, struggled when asked to make confidence judgments about their own performance in a game of Pictionary – a task relying more heavily on epistemic (i.e., "How confused am I when I look at this drawing?") rather than aleatory uncertainty. Participants made both prospective (how well they will perform in the future) and retrospective (how well they performed in the past) confidence judgments. This distinction is critical because, relative to prospective judgments, retrospective judgments rely more heavily on *mnemonic* cues (i.e., how hard did the task feel). As one might predict from the cue utilization model (Koriat, 1997), ChatGPT and Gemini's retrospective judgments were less accurate than their prospective judgments, while humans' retrospective judgments were more accurate than their prospective judgments.

These results effectively demonstrate that generative chatbots are as good as humans at making metacognitive judgments that rely on their ability to interpret information (i.e., *intrinsic* cues) to reduce aleatory uncertainty, but not particularly capable of updating these judgments based on experiential factors (i.e., *mnemonic* cues) that reduce epistemic uncertainty. Humans, on the other hand, are adept at interpreting *mnemonic* cues. It is worth noting, however, that the participants in Cash et al.'s (2024) studies were Prolific participants, not experts. As such, it is unclear how the accuracy of the generative chatbots' metacognitive judgments would stack up to

those of human experts (although if anything, one would expect experts to do better than novices, and thus also outperform the generative chatbots).

One additional challenge in assessing generative chatbots' metacognitive judgments is that they only provide these judgments when explicitly asked to do so (as in Cash et al., 2024). This is in stark contrast to humans, who have constant metacognitive experiences, even when they are not consciously attempting to engage in metacognition (Oppenheimer, 2008). However, computer scientists are currently working to design large language models (on which generative chatbots are built) that implement metacognitive strategies, with or without human prompting (Tan et al., 2024). As such, it is plausible to believe that spontaneous metacognition is right around the corner for generative chatbots, which could lessen the (meta-)cognitive gap between generative chatbots and human experts.

Conclusion

It is clear that generative chatbots and human experts approach cognition and metacognition in very different ways. Human experts develop knowledge over time, structure the knowledge into schemata, and develop intuitive understandings about how ideas within their domain of expertise are connected, allowing them to make sense of complex information, solve novel problems, and easily answer simple questions. Human experts then rely on a variety of metacognitive cues to recognize when they are likely to have made errors.

In contrast, generative chatbots are trained on a pre-determined set of inputs and structure their knowledge of the world based on this training set. As such, their knowledge is purely associative with no room for intuition. The rigid nature of generative chatbots' knowledge makes them prone to errors that human experts would not make, and they lack many of the metacognitive cues that

human experts use to detect errors when they happen. Given these limitations, it is hard to consider generative chatbots experts in nearly any domain. However, generative chatbots are still in their infancy – and may someday be telling us psychologists that we lack the cognitive and metacognitive skills to be considered experts.

Research Transparency and Openness:

*Data, Materials, and Code**:*** No data, materials, or code were used.

Funding: No funding was received to support this manuscript.

Conflict of Interest: The authors declare no conflicts of interest.

References

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review*, *19*(6), 1187–1192. https://doi.org/10.3758/s13423-012-0305-z

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 391–412. https://doi.org/10.1037/0278-7393.18.2.391

Alter, A. L., & Oppenheimer, D. M. (2008). Easy on the mind, easy on the wallet: The roles of familiarity and processing fluency in valuation judgments. *Psychonomic Bulletin & Review*, *15*(5), 985–990. https://doi.org/10.3758/PBR.15.5.985

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Burton, O. R., Bodner, G. E., Williamson, P., & Arnold, M. M. (2023). How accurate and predictive are judgments of solvability? Explorations in a two-phase anagram solving paradigm. *Metacognition and Learning*, *18*(1), 1–35. https://doi.org/10.1007/s11409-022-09313-y

Cash, T. N., & Oppenheimer, D. M. (2024). Parental rights or parental wrongs: Parents' metacognitive knowledge of the factors that influence their school choice decisions. *PLOS ONE, 19*(4), e0301768. https://doi.org/10.1371/journal.pone.0301768

Cash, T. N., Oppenheimer, D. M., & Christie, S. (2024). *Quantifying uncertAInty: Testing the accuracy of LLMs' confidence judgments*. PsyArXiv. https://doi.org/10.31234/osf.io/47df5

Cheng, V., & Yu, Z. (2023). Analyzing ChatGPT's mathematical deficiencies: Insights and contributions. In J.-L. Wu & M.-H. Su (Eds.), *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing* (pp. 188–193). The Association for Computational Linguistics and Chinese Language Processing. https://aclanthology.org/2023.rocling-1.22

Christopher, G. M., & Müller, S. (2014). Transfer of expert visual anticipation to a similar domain. *Quarterly Journal of Experimental Psychology*, *67*(1), 186–196. https://doi.org/10.1080/17470218.2013.798003

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

De Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2007). Improving metacomprehension accuracy and self-regulation in cognitive skill acquisition: The effect of learner expertise. *European Journal of Cognitive Psychology*, *19*(4–5), 671–688. https://doi.org/10.1080/09541440701326204

Deng, G., Ou, H., Liu, Y., Zhang, J., Zhang, T., & Liu, Y. (2024). *Oedipus: LLM-enhanced reasoning CAPTCHA solver*. arXiv. https://doi.org/10.48550/ARXIV.2405.07496

Derry, S. J. (1996). Cognitive schema theory in the constructivist debate. *Educational Psychologist*, *31*(3–4), 163–174. https://doi.org/10.1080/00461520.1996.9653264

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*(5), 395–416. https://doi.org/10.1037/0033-295X.85.5.395

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037/0033-295X.100.3.363

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*(4), 567–586. https://doi.org/10.1007/s10648-015-9313-7

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Keren, G. Kirkebøen, & H. Montgomery (Eds.), *Perspectives on thinking, judging, and decision making.* (pp. 21–35). Universitetsforlaget.

Frieder, S., Pinchetti, L., and Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 27699–27744). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3666122.3667327

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Grabner, R. H., Neubauer, A. C., & Stern, E. (2006). Superior performance and neural efficiency: The impact of intelligence and expertise. *Brain Research Bulletin*, *69*(4), 422–439. https://doi.org/10.1016/j.brainresbull.2006.02.009

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Han, Y., & Dunning, D. (2024). Metaknowledge of experts versus nonexperts: Do experts know better what they do and do not know? *Journal of Behavioral Decision Making*, *37*(2), e2375. https://doi.org/10.1002/bdm.2375

Huet, N., & Mariné, C. (2005). Clustering and expertise in a recall task: The effect of item organization criteria. *Learning and Instruction*, *15*(4), 297–311. https://doi.org/10.1016/j.learninstruc.2005.07.005

Imundo, M. N., Watanabe, M., Potter, A. H., Gong, J., Arner, T., & McNamara, D. S. (2024). Expert thinking with generative chatbots. *Journal of Applied Research in Memory and Cognition*.

Jiang, Y., & Hong, J. (2014). It feels fluent, but not right: The interactive effect of expected and experienced processing fluency on evaluative judgment. *Journal of Experimental Social Psychology*, *54*, 147–152. https://doi.org/10.1016/j.jesp.2014.05.004

Kaplan, I. T., & Carvellas, T. (1968). Effect of word length on anagram solution time. *Journal of Verbal Learning and Verbal Behavior*, *7*(1), 201–206. https://doi.org/10.1016/S0022-5371(68)80189-6

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183. https://doi.org/10.1016/0030-5073(77)90001-0

Long, R. (2023). Introspective capabilities in large language models. *Journal of Consciousness Studies*, *30*(9–10), 143–153. https://doi.org/10.53765/20512201.30.9.143

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*(2), 179–191. https://doi.org/10.1016/j.obhdp.2008.02.007

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–241. https://doi.org/10.1016/j.tics.2008.02.014

Patel, V. L., Cohen, T., Murarka, T., Olsen, J., Kagita, S., Myneni, S., Buchman, T., & Ghaemmaghami, V. (2011). Recovery at the edge of error: Debunking the myth of the infallible expert. *Journal of Biomedical Informatics*, *44*(3), 413–424. https://doi.org/10.1016/j.jbi.2010.09.005

Rawson, K. A., & Van Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, *58*(3), 646–668. https://doi.org/10.1016/j.jml.2007.08.004

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338–342. https://doi.org/10.1006/ccog.1999.0386

Rodriguez, C., & Oppenheimer, D. M. (2024). Creating a Bot-tleneck for malicious AI: Psychological methods for bot detection. *Behavior Research Methods*, *56*(6), 6258–6275. https://doi.org/10.3758/s13428-024-02357-9

Roese, N. J., & Sherman, J. W. (2007). Expectancy. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: A handbook of basic principles.* (Vol. 2, pp. 91–115). Guilford Press.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the Acquisition of Knowledge* (1st ed., pp. 99–135). Routledge. https://doi.org/10.4324/9781315271644-10

Salas, E., Rosen, M. A., & DiazGranados, D. (2010). Expertise-based intuition and decision making in organizations. *Journal of Management*, *36*(4), 941–973. https://doi.org/10.1177/0149206309350084

Samuels, S. J., & Flor, R. F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly*, *13*(2), 107–121. https://doi.org/10.1080/1057356970130202

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*(4), 332–348. https://doi.org/10.1207/s15327663jcp1404_2

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica*, *81*(1), 75–86. https://doi.org/10.1016/0001-6918(92)90012-3

Sher, S. (2023). *On artifice and intelligence*. Medium. https://doi.org/10.1017/CBO9780511816796.004

Simonsohn, U., Nelson, L., & Simmons, J. (2024). *Data Colada*. https://datacolada.org/about

Tan, Z., Peng, J., Chen, T., & Liu, H. (2024). *Tuning-free accountable intervention for LLM deployment—A metacognitive approach*. arXiv. http://arxiv.org/abs/2403.05636

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–251. https://doi.org/10.1016/j.cognition.2012.09.012

Van De Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, *22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica*, *98*(2–3), 141–165. https://doi.org/10.1016/S0001-6918(97)00040-1

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416