*Judgment and Decision Making* (2025), **vol. 20**:e23 1–17
doi:10.1017/jdm.2025.7

EMPIRICAL ARTICLE

# Artificial intelligence and dichotomania

Blakeley B. McShane[1], David Gal[2], and Adam Duhachek[2]

[1]Kellogg School of Management, Northwestern University, Evanston, IL, USA and [2]College of Business Administration, University of Illinois Chicago, Chicago, IL, USA

**Corresponding author:** David Gal; Email: davidgal@uic.edu

**Received:** 20 November 2024; **Revised:** 12 February 2025; **Accepted:** 12 February 2025

**Keywords:** artificial intelligence; large language models; null hypothesis significance testing; *P*-value; sociology of science; statistical significance

**Abstract**

Large language models (LLMs) such as ChatGPT, Gemini, and Claude are increasingly being used in aid or place of human judgment and decision making. Indeed, academic researchers are increasingly using LLMs as a research tool. In this paper, we examine whether LLMs, like academic researchers, fall prey to a particularly common human error in interpreting statistical results, namely 'dichotomania' that results from the dichotomization of statistical results into the categories 'statistically significant' and 'statistically nonsignificant'. We find that ChatGPT, Gemini, and Claude fall prey to dichotomania at the 0.05 and 0.10 thresholds commonly used to declare 'statistical significance'. In addition, prompt engineering with principles taken from an American Statistical Association *Statement on Statistical Significance and* P-*values* intended as a corrective to human errors does not mitigate this and arguably exacerbates it. Further, more recent and larger versions of these models do not necessarily perform better. Finally, these models sometimes provide interpretations that are not only incorrect but also highly erratic.

## 1. Introduction

Artificial intelligence models—and large language models (LLMs) such as OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude in particular—are increasingly being used in aid or place of human judgment and decision making. For example, LLMs have been used for medical diagnosis (Meng et al., 2024), legal contracts (Martin et al., 2024), recommendation systems (Zhao et al., 2024), time series forecasting (Tang et al., 2025), and a host of other applications.

LLMs produce text (e.g., in response to queries about medicine or law) by estimating the probability that a given token (i.e., a character, word, or subword such as a prefix or suffix) or sequence of tokens would appear within some larger sequence of tokens and then randomly choosing one from those that have high estimated probability. A key characteristic that distinguishes LLMs from less sophisticated language models such as autocomplete or predictive text is their scale: LLMs involve hundreds of billions or trillions of parameters that are estimated on the basis of a vast amount of human-generated text including from sources such as books, research papers, and the internet. Given this, LLMs produce text that is similar in nature to human-generated text (indeed, so similar that it is difficult for humans to distinguish from human-generated text (Jones and Bergen, 2024)).

One consequence of this similarity is that errors and biases found in human reasoning are also found in LLM-generated text. For example, when asked to provide confidence judgments about predictions,

LLMs show errors and biases similar to those of human confidence judgments (Cash et al., 2024). LLMs also show a status quo bias similar to that of humans (Horton, 2023). In addition, LLMs show a range of cognitive biases that are similar to or even exceed those of humans (Koo et al., 2024).

Nonetheless, users of LLMs use them because they seek and believe them to provide accurate responses. This is not necessarily unreasonable because LLMs perform strongly when benchmarked for accuracy against objective standards. Indeed, leaderboards (see, e.g., Kirkovska, 2024) benchmark and compare the performance of LLMs on tasks related to commonsense reasoning; fluid intelligence; reading comprehension; knowledge of the humanities, social sciences, and hard sciences, and other areas; mathematics competition problems; and coding. Further, developers of successive versions of LLMs claim improvement over prior versions as well as over other LLMs by showing superior performance on such benchmarks.

Academic researchers are increasingly using LLMs as a research tool. For example, LLMs have been used to summarize research papers (Cai et al., 2024; Jin et al., 2024), provide feedback on them (Liang et al., 2024), conduct literature reviews (Antu et al., 2023), and provide structured science summarizations (Nechakhin et al., 2024). Toward this end, OpenAI recently (February 2, 2025) released Deep Research, a new capability for ChaptGPT that can 'find, analyze, and synthesize hundreds of online sources to create a comprehensive report at the level of a research analyst' (OpenAI, 2025b). LLMs have also been used to teach statistical analysis (Xing, 2024) and for automated statistical analysis (Jansen et al., 2023). They have even been used for automated scientific discovery, that is, to produce research by generating ideas, writing code, running experiments, visualizing results, and ultimately writing and evaluating a research paper (Lu et al., 2024; see also Gans, 2025a, 2025b).

It would seem that a necessary precondition for performing such research tasks competently would be the ability to accurately interpret statistical results. However, human interpretations of statistical results—specifically, those made by academic researchers—are prone to error. Indeed, in a corrective measure, the Board of Directors of the American Statistical Association (ASA) issued a *Statement on Statistical Significance and* P*-values* (Wasserstein and Lazar, 2016) that noted that 'statistical significance' and *P*-values are 'commonly misused and misinterpreted' and offered six principles regarding their 'proper use and interpretation'. Principle 3 of the ASA Statement ('Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold') warns against a particularly common human error, namely 'dichotomania' that results from the dichotomization of statistical results into the categories 'statistically significant' and 'statistically nonsignificant' (Amrhein et al., 2019; Greenland, 2017).

Given that errors found in human reasoning are also found in LLM-generated text due to the fact that LLMs are estimated on human-generated text, one might hazard that LLMs too fall prey to dichotomania. Alternatively, given their strong performance when benchmarked for accuracy against objective standards, perhaps LLMs are more resistant—or via prompt engineering could be made to be more resistant—to dichotomania than humans are. This is an important matter to address given that LLMs are increasingly being used by academic researchers as a research tool and that a necessary precondition for performing research tasks competently is the ability to accurately interpret statistical results.

We find that ChatGPT, Gemini, and Claude fall prey to dichotomania at the 0.05 and 0.10 thresholds commonly used to declare 'statistical significance'. In addition, prompt engineering with either the six principles of the ASA Statement or Principle 3 in particular does not mitigate this and arguably exacerbates it. Further, more recent and larger versions of these models do not necessarily perform better. Finally, these models sometimes provide interpretations that are not only incorrect but also highly erratic.

In the remainder of this paper, we provide a brief literature review. We next discuss our methods, specifically the three questions we use to assess dichotomania, the prompt engineering that we employ in an attempt to mitigate it, and some implementation details. We then discuss the results for each question in turn. We finally conclude with a brief discussion.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

## 2. Literature review

### *2.1. Overview*

Human interpretations of statistical results—specifically, those made by academic researchers—are prone to error. The ASA Statement notes that *P*-values and 'statistical significance' are commonly misinterpreted.

Regarding the former, researchers misinterpret the *P*-value as, among other things, the probability that some target hypothesis is true, one minus the probability that some alternative hypothesis is true, and one minus the probability of replication. For example, Gigerenzer (2004) reports on a study conducted by Haller and Krauss (2002) on psychology professors, lecturers, teaching assistants, and students (see also Oakes (1986)). Subjects were given the result of a simple *t*-test of two independent means ($t = 2.7$, $df = 18$, $p = 0.01$) and were asked six true or false questions about the result that were designed to test common misinterpretations of the *P*-value. All six of the statements were false and, despite the fact that the study materials noted that 'several or none of the statements may be correct', (i) none of the forty-four students, (ii) only four of the thirty-nine professors and lectures who did not teach statistics, and (iii) only six of the thirty professors and lectures who did teach statistics marked all as false. For a review of related studies, see Gigerenzer (2018); see also Goodman (2008) and Greenland et al. (2016).

Regarding the latter, a particularly common error is dichotomania that results from the dichotomization of statistical results into the categories 'statistically significant' and 'statistically nonsignificant'. As per Principle 3 of the ASA Statement, researchers wrongly interpret results that are 'statistically significant' as demonstrating an effect and those that are 'statistically nonsignificant' as demonstrating no effect. This leads to a further error, namely wrongly interpreting a result that is 'statistically significant' and another result that is 'statistically nonsignificant' as being in conflict with one another. That is, categorizing these results differently induces researchers to draw the incorrect conclusion that the results thus categorized are categorically different. Instead, as Gelman and Stern (2006) point out, the difference between a result that is 'statistical significant' and another result that is 'statistical nonsignificant' is itself often not 'statistically significant' (i.e., because *P*-values naturally vary a great deal from sample to sample, and thus sampling variation alone can easily cause large differences in *P*-values—not only *P*-values that fall just barely to either side of some threshold; Greenland, 2019; McShane et al., 2024). In addition, researchers wrongly believe that 'statistical significance' indicates practical importance (Boring, 1919; Freeman, 1993). They also believe 'statistical significance' supports attributions of causality (Holman et al., 2001).

### *2.2. McShane and Gal*

McShane and Gal (2016, 2017) report studies of dichotomania in academic researchers, in particular, in applied researchers across a wide variety of fields including medicine, epidemiology, cognitive science, psychology, business, and economics as well as in statisticians. Because we use the three questions they used to assess dichotomania in researchers to assess dichotomania in LLMs, we here briefly review their studies and results.

Their studies presented researchers with a summary of a hypothetical experiment comparing two treatments in which the *P*-value for the comparison was manipulated to be 'statistically significant' or 'statistically nonsignificant'. The researchers were then asked about descriptions of the data presented in the summary or to make predictions and decisions on the basis of it.

Their results showed that applied researchers fall prey to dichotomania. Specifically, they interpret the *P*-value dichotomously rather than continuously, focusing on whether or not it is below the 0.05 threshold rather than its magnitude. Further, they fixate on the *P*-value even when it is irrelevant, for example, when asked about descriptive statistics. In addition, they ignore other information including the magnitude of the treatment difference. Their results also showed that statisticians also fall prey to dichotomania although to a somewhat lesser degree than applied researchers.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

## 3. Methods

### *3.1. Question 1: Description*

The first question we use to assess dichotomania is the McShane and Gal (2016, 2017) question about description. In particular, we ask:

> Below is a summary of a study from an academic paper.
>
> The study aimed to test how different interventions might affect terminal cancer patients' survival. Subjects were randomly assigned to one of two groups. Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure. Subjects were then tracked until all had died. Subjects in Group A lived, on average, 8.2 months post-diagnosis whereas subjects in Group B lived, on average, 7.5 months post-diagnosis ($p = 0.01$).
>
> Which statement is the most accurate summary of the results?
>
> A. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the subjects who were in Group A was *greater* than that lived by the subjects who were in Group B.
> B. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the subjects who were in Group A was *less* than that lived by the subjects who were in Group B.
> C. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the subjects who were in Group A was *no different* than that lived by the subjects who were in Group B.
> D. Speaking only of the subjects who took part in this particular study, it *cannot be determined* whether the average number of post-diagnosis months lived by the subjects who were in Group A was greater/no different/less than that lived by the subjects who were in Group B.

In asking this question, we vary the observed *P*-value $p$ presented in the question across all values used in prior research: 0.01, 0.025, 0.075, 0.125, 0.175, and 0.27. Most importantly for our purposes, we also use the values of 0.049 and 0.051 as well as 0.099 and 0.101 to assess dichotomania at the respective 0.05 and 0.10 thresholds commonly used to declare 'statistical significance'.

In addition to the version of the question presented above, we followed prior research in using a version that omits the response option preamble 'Speaking only of the subjects who took part in this particular study' from each of the four response options.

The correct answer to this question is Option A regardless of the observed *P*-value presented and the presence or absence of the response option preamble: all four response options are descriptive statements and indeed the average number of post-diagnosis months lived by the subjects who were in Group A was greater than that lived by the subjects who were in Group B (i.e., $8.2 > 7.5$).

Nonetheless, in prior research, academic researchers were much more likely to answer the question correctly when the observed *P*-value presented was below 0.05. Further, the presence or absence of the response option preamble did not substantially affect the pattern of results.

### *3.2. Question 2: Prediction*

The second question we use to assess dichotomania is the McShane and Gal (2016, 2017) question about prediction. In particular, we ask:

> Below is a summary of a study from an academic paper:
>
> The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population and then randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of subjects who took Drug A recovered

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

from the disease while forty-four percent (44%) of subjects who took Drug B recovered from the disease.

A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a *p*-value of 0.01.

Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?

A. A person drawn randomly from the same population as the subjects in the study is *more likely* to recover from the disease if given Drug A than if given Drug B.
B. A person drawn randomly from the same population as the subjects in the study is *less likely* to recover from the disease if given Drug A than if given Drug B.
C. A person drawn randomly from the same population as the subjects in the study is *equally likely* to recover from the disease if given Drug A than if given Drug B.
D. It *cannot be determined* whether a person drawn randomly from the same population as the subjects in the study is more/less/equally likely to recover from the disease if given Drug A or if given Drug B.

In asking this question, we again vary the observed *P*-value *p* presented in the question across all values used in prior research: 0.01, 0.025, 0.075, 0.125, 0.175, and 0.26. Most importantly for our purposes, we also use the values of 0.049 and 0.051 as well as 0.099 and 0.101 to assess dichotomania at the respective 0.05 and 0.10 thresholds commonly used to declare 'statistical significance'.

The correct answer to this question clearly depends on whether or not Drug A is more effective than Drug B. The observed *P*-value is one measure of the strength of the evidence regarding this. However, and congruent with Principle 3 of the ASA Statement, the correct answer 'should not be based only on whether a *P*-value passes a specific threshold'. Further, from a Bayesian perspective, the correct answer is Option A regardless of the observed *P*-value presented. Indeed, under the non-informative prior encouraged by the question wording, the probability that Drug A is more effective than Drug B is a decreasing linear function of the observed *P*-value (i.e., it is one minus the two-sided observed *P*-value divided by two).

Nonetheless, in prior research, academic researchers were much more likely to answer the question correctly when the observed *P*-value presented was below 0.05.

### *3.3. Question 3: Decision*

The third question we use to assess dichotomania is the McShane and Gal (2016, 2017) question about decision-making. In particular, we present the same study summary as in Question 2 but ask:

Assuming no prior studies have been conducted with these drugs, what drug would you advise physicians treating patients from the same population as those in the study prescribe for their patients?

A. I would advise Drug A.
B. I would advise Drug B.
C. I would advise that there is no difference between Drug A and Drug B.

In asking this question, we again vary the observed *P*-value *p* presented in the question across all values used in prior research: 0.01, 0.025, 0.075, 0.125, 0.175, and 0.26. Most importantly for our purposes, we also use the values of 0.049 and 0.051 as well as 0.099 and 0.101 to assess dichotomania at the respective 0.05 and 0.10 thresholds commonly used to declare 'statistical significance'.

Like Question 2, the correct answer to this question clearly depends on whether or not Drug A is more effective than Drug B, 'should not be based only on whether a *P*-value passes a specific threshold', and from a Bayesian perspective is Option A regardless of the observed *P*-value presented.

Nonetheless, in prior research, academic researchers were much more likely to answer the question correctly when the observed *P*-value presented was below 0.05. Further, they were much more likely

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

to answer this question correctly as compared to Question 2 when the observed *P*-value presented was above 0.05 suggesting that shifting the question from one about prediction to one about decision-making reduces but does not eliminate dichotomania.

### *3.4. Prompt engineering*

In addition to using the versions of the questions presented above used in prior research, we also use two additional versions that employ prompt engineering in an attempt to mitigate dichotomania.

The second version of each question adds the six principles of the ASA Statement after presenting the response options. Specifically, it adds:

> In considering this question, it might be helpful to bear in mind the six principles articulated in the 2016 American Statistical Association *Statement on Statistical Significance and* P-*values*:
>
> Principle 1. *P*-values can indicate how incompatible the data are with a specified statistical model.
>
> Principle 2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
>
> Principle 3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
>
> Principle 4. Proper inference requires full reporting and transparency.
>
> Principle 5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
>
> Principle 6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

The third version of each question adds only the most relevant principle, namely Principle 3, of the ASA Statement after presenting the response options. Specifically, it adds:

> In considering this question, it might be helpful to bear in mind a principle articulated in the 2016 American Statistical Association *Statement on Statistical Significance and* P-*values*: Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

### *3.5. Implementation details*

We assess dichotomania in ChatGPT, Gemini, and Claude. Specifically, and in decreasing order of model recency and size within each family of models, we assess (i) ChatGPT 4o, ChatGPT 4o Mini, ChatGPT 4o Turbo, ChatGPT 4, and ChatGPT 3.5 Turbo, (ii) Gemini 1.5 Pro, Gemini 1.5 Flash, and Gemini 1.0 Pro, and (iii) Claude 3.5 Sonnet, Claude 3 Opus, Claude 3 Sonnet, and Claude 3 Haiku.

Because LLM responses are stochastic, we assess dichotomania across 100 independent trials for each model version, question version, and observed *P*-value. We do so via the Application Programming Interface. We leave all hyperparameters (e.g., temperature, top probability mass, maximum number of tokens) at their default settings. However, we explicitly set temperature to the 'neutral' default (i.e., to one for ChatGPT and Gemini and to one-half for Claude because temperature ranges from zero to two for the former and zero to one for the latter). We conducted our trials in August 2024 (ChatGPT) and September 2024 (Gemini and Claude).

Due to the large number of trials in total, we attempt to encourage responses that facilitate automatic processing. For example, we at the end of the prompt add:

> Please answer only Option A, Option B, Option C, or Option D.

We also explicitly labeled the response options as Option A, Option B, Option C, and Option D.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

## 4. Results

### *4.1. Question 1: Description*

In presenting our results, we focus on the number of times that the response was Option A (i.e., the correct answer to this question) for each model version, question version, and observed *P*-value across the 100 independent trials. We also focus on the observed *P*-values of 0.01 (to establish a baseline) as well as 0.049 versus 0.051 and 0.099 versus 0.101 (to assess dichotomania at the respective 0.05 and 0.10 thresholds). Full results can be found in our Supplemental Materials.

The results for Question 1 can be found in Table 1. When the question is presented with the response option preamble 'Speaking only of the subjects who took part in this particular study' (i.e., Table 1(a)–1(c)), model performance is reasonably strong. All versions of ChatGPT except ChatGPT 3.5 Turbo, the oldest and smallest ChatGPT model version, nearly always choose Option A. However, ChatGPT 3.5 Turbo performs erratically as the observed *P*-value increases from 0.051 to 0.099 and dichotomously at the 0.05 and 0.10 thresholds. Prompt engineering does not mitigate this and arguably exacerbates it.

Gemini 1.5 Pro performs dichotomously at the 0.10 threshold when prompted with the six principles of the ASA Statement. Gemini 1.5 Flash performs erratically as the observed *P*-value increases from 0.051 to 0.099 and dichotomously (though in the opposite of the expected direction) at the 0.10 threshold. Prompt engineering again arguably exacerbates rather than mitigates this behavior. Gemini 1.0 Pro, the oldest and smallest Gemini model version, performs well and best of all of the Gemini model versions.

Finally, Claude 3.5 Sonnet and Claude 3 Haiku perform well but Claude 3 Opus and Claude 3 Sonnet do not. Claude 3 Opus performs dichotomously at the 0.05 threshold and dichotomously (though in the opposite of the expected direction) at the 0.10 threshold. It also performs erratically (though in the opposite of the expected direction) as the observed *P*-value increases from 0.051 to 0.099. Prompt engineering does not mitigate this behavior. Claude 3 Sonnet performs dichotomously at the 0.10 threshold and erratically as the observed *P*-value increases from 0.051 to 0.099 when prompted with Principle 3 of the ASA Statement.

When the question is presented without the response option preamble (i.e., Table 1(d)–1(f)), model performance is considerably worse. ChatGPT 4o performs dichotomously at the 0.05 and especially the 0.10 thresholds (and erratically as the observed *P*-value increases from 0.051 to 0.099 when prompted with Principle 3 of the ASA Statement). ChatGPT 4 Turbo also performs dichotomously at the 0.10 threshold. ChatGPT 3.5 Turbo performance is poor and erratic.

Gemini 1.5 Pro performs dichotomously at the 0.05 threshold when prompted with either the six principles of the ASA Statement or Principle 3 and at the 0.10 threshold when not. It also performs erratically as the observed *P*-value increases from 0.051 to 0.099 when not prompted. Gemini 1.5 Flash performs erratically as the observed *P*-value increases from 0.051 to 0.099 and dichotomously (though in the opposite of the expected direction) at the 0.10 threshold when prompted with Principle 3 of the ASA Statement. Gemini 1.0 Pro performs well and best of all of the Gemini model versions.

Claude 3.5 Sonnet and Claude 3 Opus perform dichotomously at the 0.05 threshold. Claude 3 Sonnet performs erratically as the observed *P*-value increases from 0.051 to 0.099 and dichotomously at the 0.10 threshold. Claude 3 Haiku, the oldest and smallest Claude model version, performs well and best of all of the Claude model versions.

In addition to the incorrect and erratic performance noted above, we discuss some additional problems aggregating across all question versions and observed *P*-values we considered. ChatGPT 4o chose Option C (i.e., that the two groups were no different) on 110 trials presented without the response option preamble, and both ChatGPT 4o Mini and ChatGPT 4 did so on 1 trial. ChatGPT 4 Turbo chose Option C on 11 trials presented with the response option preamble and 588 presented without. ChatGPT 3.5 Turbo chose Option C on 57 trials presented with the response option preamble and 242 presented without. This is problematic because Option C exemplifies the 'proof of the null' fallacy that has been lamented for over a century (Fisher, 1935; Pearson, 1906). Further, ChatGPT 3.5 Turbo also

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

***Table 1.*** *Question 1: Description results.*

| Model version | Prompt engineering | Observed *P*-value 0.01 | 0.049 | 0.051 | 0.099 | 0.101 |
|---|---|---|---|---|---|---|
| (a) ChatGPT with response option preamble. | | | | | | |
| ChatGPT 4o | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 99 | 100 |
| ChatGPT 4o Mini | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| ChatGPT 4 Turbo | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| ChatGPT 4 | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| ChatGPT 3.5 Turbo | None | 100 | 100 | 98 | 51 | 23 |
| | ASA | 100 | 100 | 80 | 23 | 8 |
| | ASA P3 | 100 | 99 | 82 | 17 | 0 |
| (b) Gemini with response option preamble. | | | | | | |
| Gemini 1.5 Pro | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 70 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| Gemini 1.5 Flash | None | 100 | 100 | 100 | 59 | 100 |
| | ASA | 100 | 100 | 100 | 0 | 93 |
| | ASA P3 | 100 | 100 | 100 | 7 | 55 |
| Gemini 1.0 Pro | None | 100 | 99 | 100 | 98 | 99 |
| | ASA | 99$^\star$ | 100 | 99 | 97 | 92 |
| | ASA P3 | 100 | 100 | 100 | 100 | 97 |
| (c) Claude with response option preamble. | | | | | | |
| Claude 3.5 Sonnet | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| Claude 3 Opus | None | 100 | 100 | 41 | 56 | 72 |
| | ASA | 100 | 100 | 54 | 97 | 100 |
| | ASA P3 | 100 | 100 | 36 | 70 | 95 |
| Claude 3 Sonnet | None | 100 | 100 | 100 | 100 | 74 |
| | ASA | 100 | 100 | 100 | 97 | 16 |
| | ASA P3 | 100 | 100 | 100 | 1 | 0 |
| Claude 3 Haiku | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 95 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |

(Contiuned)

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

***Table 1.*** *Continued.*

| | | | | | | |
|---|---|---|---|---|---|---|
| (d) ChatGPT without response option preamble. | | | | | | |
| ChatGPT 4o | None | 100 | 100 | 94 | 95 | 79 |
| | ASA | 100 | 100 | 98 | 99 | 85 |
| | ASA P3 | 100 | 100 | 86 | 95 | 84 |
| ChatGPT 4o Mini | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| ChatGPT 4 Turbo | None | 100 | 100 | 100 | 100 | 90 |
| | ASA | 100 | 100 | 100 | 97 | 82 |
| | ASA P3 | 100 | 100 | 100 | 90 | 63 |
| ChatGPT 4 | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 98 |
| ChatGPT 3.5 Turbo | None | 100 | 98 | 97 | 28 | 12 |
| | ASA | 100 | 100 | 95 | 40 | 14 |
| | ASA P3 | 100 | 100 | 90 | 7 | 1 |
| (e) Gemini without response option preamble. | | | | | | |
| Gemini 1.5 Pro | None | 100 | 100 | 100 | 33 | 8 |
| | ASA | 100 | 100 | 11 | 0 | 0 |
| | ASA P3 | 99 | 100 | 0 | 0 | 0 |
| Gemini 1.5 Flash | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 38 | 98 |
| Gemini 1.0 Pro | None | 100 | 100 | 100 | 99 | 99 |
| | ASA | 100 | 100 | 100 | 96 | 94 |
| | ASA P3 | 100 | 100 | 100 | 99 | 95 |
| (f) Claude without response option preamble. | | | | | | |
| Claude 3.5 Sonnet | None | 100 | 100 | 0 | 0 | 0 |
| | ASA | 100 | 100 | 0 | 0 | 0 |
| | ASA P3 | 100 | 100 | 0 | 0 | 0 |
| Claude 3 Opus | None | 100 | 100 | 2 | 0 | 0 |
| | ASA | 100 | 100 | 0 | 0 | 0 |
| | ASA P3 | 100 | 100 | 0 | 0 | 0 |
| Claude 3 Sonnet | None | 100 | 100 | 100 | 31 | 0 |
| | ASA | 100 | 100 | 100 | 19 | 0 |
| | ASA P3 | 100 | 100 | 95 | 0 | 0 |
| Claude 3 Haiku | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |

Note: Each cell of the table gives the number of times that the response was Option A for each model version, question version, and observed *P*-value across 100 independent trials. None denotes the version of the question used in prior research; ASA (ASA P3) denotes the version that employs prompt engineering with the six principles (Principle 3) of the ASA Statement. A star denotes that the given model version did not respond on one of the trials for the given question version and observed *P*-value.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

chose Option B (i.e., that the better performing group performed worse) on 158 trials presented with the response option preamble and 100 presented without. This is clearly problematic.

Gemini 1.5 Pro chose Option C on 770 trials presented without the response option preamble. Gemini 1.0 Pro chose Option C on 63 trials presented with the response option preamble and 78 presented without. Gemini 1.0 Pro also chose Option B on 9 trials presented with the response option preamble and 3 presented without.

Finally, Claude 3 Opus chose Option C on 336 trials presented with the response option preamble and 180 presented without. Claude 3 Sonnet chose Option C on 776 trials presented with the response option preamble and 1,566 presented without. Claude 3 Haiku chose Option C on 160 trials presented with the response option preamble and 141 presented without.

### *4.2. Question 2: Prediction*

The results for Question 2 can be found in Table 2. Model performance is uniformly poor. All versions of ChatGPT perform dichotomously at the 0.05 threshold and also, when applicable, at the 0.10 threshold. Prompt engineering does not mitigate this behavior.

Gemini 1.5 Pro performs dichotomously at the 0.05 threshold, erratically as the observed *P*-value increases from 0.01 to 0.049 when prompted with six principles of the ASA Statement, and inexplicably when prompted with Principle 3. Gemini 1.5 Flash performs dichotomously at the 0.05 threshold; prompt engineering does not mitigate this and arguably makes performance more erratic. Gemini 1.0 Pro performs erratically when the observed *P*-value is 0.01 and dichotomously at the 0.05 threshold.

Claude 3.5 Sonnet performs dichotomously at the 0.05 threshold and erratically as the observed *P*-value increases from 0.01 to 0.049 when prompted with Principle 3 of the ASA Statement. Claude 3 Opus perform dichotomously at the 0.05 threshold. Claude 3 Sonnet performs less dichotomously at the 0.05 threshold but dichotomization is exacerbated when prompted with six principles of the ASA Statement; Claude 3 Sonnet also performs erratically as the observed *P*-value increases from 0.051 to 0.099. Claude 3 Haiku performs well except when prompted with either the six principles of the ASA Statement or Principle 3, which causes it to perform dichotomously at the 0.05 threshold.

In addition to the incorrect and erratic performance noted above, we discuss some additional problems aggregating across all question versions and observed *P*-values we considered. ChatGPT 3.5 Turbo chose Option C on 4 trials. Gemini 1.0 Pro chose Option B on 26 trials and Option C on 168 trials. Claude 3 Opus chose Option C on 1 trial and Claude 3 Haiku chose Option C on 7 trials.

### *4.3. Question 3: Decision*

The results for Question 3 can be found in Table 3. Model performance varies considerably by model version. ChatGPT 4o performs dichotomously at the 0.10 threshold. ChatGPT 4o Mini, ChatGPT 4 Turbo, ChatGPT 4 perform well. ChatGPT 3.5 Turbo performs dichotomously at the 0.05 and 0.10 thresholds and erratically as the observed *P*-value increases from 0.051 to 0.099.

Gemini 1.5 Pro and Gemini 1.5 Flash perform dichotomously at the 0.05 threshold. Gemini 1.0 Pro performs erratically when the observed *P*-value is 0.01 and dichotomously at the 0.05 and 0.10 thresholds

Claude 3.5 Sonnet performs erratically as the observed *P*-value increases from 0.051 to 0.099. Claude 3 Opus performs erratically as the observed *P*-value increases from 0.01 to 0.049 and dichotomously at the 0.05 threshold. Claude 3 Sonnet performs dichotomously at the 0.05 threshold. Claude 3 Haiku performs well.

Prompt engineering does not much impact these results.

In addition to the incorrect and erratic performance noted above, we discuss some additional problems aggregating across all question versions and observed *P*-values we considered. ChatGPT 3.5 Turbo chose Option B on 69 trials. Gemini 1.0 Pro chose Option B on 46 trials.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

***Table 2.*** *Question 2: Prediction results.*

| Model version | Prompt engineering | Observed *P*-value 0.01 | 0.049 | 0.051 | 0.099 | 0.101 |
|---|---|---|---|---|---|---|
| (a) ChatGPT. | | | | | | |
| ChatGPT 4o | None | 100 | 100 | 31 | 10 | 0 |
| | ASA | 100 | 100 | 3 | 1 | 0 |
| | ASA P3 | 100 | 99 | 10 | 5 | 1 |
| ChatGPT 4o Mini | None | 100 | 100 | 99 | 18 | 0 |
| | ASA | 100 | 100 | 10 | 7 | 0 |
| | ASA P3 | 100 | 100 | 13 | 3 | 0 |
| ChatGPT 4 Turbo | None | 100 | 100 | 0 | 0 | 0 |
| | ASA | 100 | 97 | 0 | 0 | 0 |
| | ASA P3 | 100 | 88 | 0 | 0 | 0 |
| ChatGPT 4 | None | 100 | 100 | 1 | 0 | 0 |
| | ASA | 100 | 100 | 1 | 0 | 0 |
| | ASA P3 | 100 | 100 | 0 | 0 | 0 |
| ChatGPT 3.5 Turbo | None | 100 | 100 | 4 | 0 | 0 |
| | ASA | 99 | 100 | 2 | 0 | 0 |
| | ASA P3 | 96 | 95 | 3 | 0 | 1 |
| (b) Gemini. | | | | | | |
| Gemini 1.5 Pro | None | 100 | 100 | 0 | 0 | 0 |
| | ASA | 100 | 0 | 0 | 0 | 0 |
| | ASA P3 | 0 | 0 | 0 | 0 | 0 |
| Gemini 1.5 Flash | None | 100 | 100 | 0 | 0 | 0 |
| | ASA | 100 | 100 | 0 | 0 | 0 |
| | ASA P3 | 100 | 0 | 0 | 0 | 0 |
| Gemini 1.0 Pro | None | 85 | 75 | 23 | 10 | 7 |
| | ASA | 63 | 45⋆ | 4⋆ | 7 | 5 |
| | ASA P3 | 55 | 38 | 17 | 2 | 1 |
| (c) Claude. | | | | | | |
| Claude 3.5 Sonnet | None | 100 | 100 | 0 | 0 | 0 |
| | ASA | 100 | 100 | 0 | 0 | 0 |
| | ASA P3 | 100 | 55 | 0 | 0 | 0 |
| Claude 3 Opus | None | 100 | 100 | 0 | 0 | 0 |
| | ASA | 100 | 99 | 0 | 0 | 0 |
| | ASA P3 | 100 | 100 | 0 | 0 | 0 |
| Claude 3 Sonnet | None | 100 | 100 | 93 | 1 | 0 |
| | ASA | 100 | 100 | 43 | 0 | 0 |
| | ASA P3 | 100 | 100 | 90 | 0 | 0 |
| Claude 3 Haiku | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 0 | 0 | 0 |
| | ASA P3 | 100 | 100 | 17 | 34 | 0 |

Note: Each cell of the table gives the number of times that the response was Option A for each model version, question version, and observed *P*-value across 100 independent trials. None denotes the version of the question used in prior research; ASA (ASA P3) denotes the version that employs prompt engineering with the six principles (Principle 3) of the ASA Statement. A star denotes that the given model version did not respond on one of the trials for the given question version and observed *P*-value.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

***Table 3.*** *Question 3: Decision results.*

| Model version | Prompt engineering | Observed *P*-value 0.01 | 0.049 | 0.051 | 0.099 | 0.101 |
|---|---|---|---|---|---|---|
| (a) ChatGPT. | | | | | | |
| ChatGPT 4o | None | 100 | 100 | 99 | 92 | 36 |
| | ASA | 100 | 100 | 98 | 97 | 39 |
| | ASA P3 | 100 | 100 | 100 | 94 | 47 |
| ChatGPT 4o Mini | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| ChatGPT 4 Turbo | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 99 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| ChatGPT 4 | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 100 |
| ChatGPT 3.5 Turbo | None | 100 | 100 | 63 | 19 | 2 |
| | ASA | 100 | 100 | 62 | 12 | 2 |
| | ASA P3 | 100 | 100 | 63 | 19 | 6 |
| (b) Gemini. | | | | | | |
| Gemini 1.5 Pro | None | 100 | 100 | 7 | 0 | 0 |
| | ASA | 100 | 100 | 6 | 0 | 0 |
| | ASA P3 | 100 | 100 | 4 | 0 | 0 |
| Gemini 1.5 Flash | None | 100 | 100 | 0 | 6 | 0 |
| | ASA | 100 | 100 | 0 | 3 | 0 |
| | ASA P3 | 100 | 100 | 0 | 5 | 0 |
| Gemini 1.0 Pro | None | 88$^{\star\star}$ | 71 | 25 | 16 | 14 |
| | ASA | 83 | 68 | 27 | 24$^{\star\star}$ | 19 |
| | ASA P3 | 78$^{\star}$ | 65 | 27$^{\star\star}$ | 16 | 15$^{\star}$ |
| (c) Claude. | | | | | | |
| Claude 3.5 Sonnet | None | 100 | 100 | 99 | 0 | 0 |
| | ASA | 100 | 100 | 98 | 0 | 0 |
| | ASA P3 | 100 | 100 | 97 | 0 | 0 |
| Claude 3 Opus | None | 100 | 80 | 0 | 0 | 0 |
| | ASA | 100 | 73 | 0 | 0 | 0 |
| | ASA P3 | 100 | 72 | 0 | 0 | 0 |
| Claude 3 Sonnet | None | 100 | 100 | 1 | 0 | 0 |
| | ASA | 100 | 100 | 2 | 0 | 0 |
| | ASA P3 | 100 | 100 | 1 | 1 | 0 |
| Claude 3 Haiku | None | 100 | 100 | 100 | 100 | 100 |
| | ASA | 100 | 100 | 100 | 100 | 100 |
| | ASA P3 | 100 | 100 | 100 | 100 | 99 |

Note: Each cell of the table gives the number of times that the response was Option A for each model version, question version, and observed *P*-value across 100 independent trials. None denotes the version of the question used in prior research; ASA (ASA P3) denotes the version that employs prompt engineering with the six principles (Principle 3) of the ASA Statement. A star denotes (two stars denote) that the given model version did not respond on one (two) of the trials for the given question version and observed *P*-value.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

## 5. Discussion

We find that ChatGPT, Gemini, and Claude fall prey to dichotomania at the 0.05 and 0.10 thresholds commonly used to declare 'statistical significance'. In addition, prompt engineering with either the six principles of the ASA Statement or Principle 3 in particular does not mitigate this and arguably exacerbates it. Further, more recent and larger versions of these models do not necessarily perform better. Finally, these models sometimes provide interpretations that are not only incorrect but also highly erratic.

The fact that more recent and larger versions of these models do not necessarily perform better is not particularly surprising when one considers that (i) LLMs seem to 'solve' questions by 'learning' specific mappings from the estimation data rather by 'reasoning' or developing an 'understanding' of the question (Kapoor et al., 2024), (ii) human interpretations of statistical results serve as the estimation data for LLMs, and (iii) human interpretations of statistical results are prone to error. Consequently, there is reason to doubt that current approaches to improving the performance of LLMs—namely by increasing the size of the model, the amount of estimation data, and the cost of computation and by providing human feedback—will lead to better interpretations of statistical results by future versions of these models.

In fact, there is arguably reason for pessimism. Specifically, although more recent and larger versions of these models do indeed perform better than less recent and smaller versions on difficult questions, they do not necessarily perform better on simple questions; in addition, models that successfully solve difficult questions also fail at simple questions (Zhou et al., 2024). These results are consistent with our finding that more recent and larger versions of these models do not necessarily perform better at interpreting statistical results and casts doubt on when if ever they will perform well.

This raises the question of how to improve the performance of LLMs. While we were not particularly surprised to find that LLMs fall prey to dichotomania, we had expected that prompt engineering with the six principles of the ASA Statement and with Principle 3 in particular would mitigate it given that (i) the ASA Statement was issued as a corrective to common misuses and misinterpretations of 'statistical significance' and *P*-values, (ii) the six principles contained within it were offered to promote their proper use and interpretation, (iii) and Principle 3 warns against dichotomania in particular. We were surprised that it does not and that it arguably exacerbates it.

We were also surprised that the presence versus absence of the response option preamble in Question 1 did seem to improve performance because it did not substantially affect that of academic researchers in prior research (McShane and Gal, 2016, 2017). This suggests that prompt engineering at least in some form or another does have the potential to improve performance.

Alternative approaches to prompt engineering include (i) few-shot prompting that provides one or more exemplar questions along with the answers to those questions prior to the focal question as opposed to our current zero-shot prompting that proceeds immediately to the focal question, (ii) zero-shot chain-of-thought prompting that includes a phrase such as 'Let's think step by step' after the focal question, and (iii) few-shot chain-of-thought prompting that provides not only one or more exemplar questions and answers but also the reasoning behind those answers.

While these approaches have improved performance in other contexts (Kojima et al., 2023; Yu et al., 2023), we are not particularly optimistic that they would in our context for three reasons. First, we believe that the prompt engineering that we employed would function similarly to these approaches and in particular to chain-of-thought prompting. Specifically, we reason that prompt engineering with the six principles of the ASA Statement and with Principle 3 in particular would provide exactly the type of reasoning to be used for the three reasons given three paragraphs prior. Nonetheless, prompt engineering with either the six principles of the ASA Statement or Principle 3 in particular does not mitigate and arguably exacerbates dichotomania.

Although we cannot definitively ascertain why this is the case, perhaps the fact that the six principles of the ASA Statement focus on *P*-values and 'statistical significance' causes LLMs to give *P*-values and 'statistical significance' additional weight. In addition, perhaps the fact that Principle 3 of the

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

ASA Statement ('Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold') focuses on whether a *P*-value passes a specific threshold or the presence of the word 'only' in it causes LLMs to give whether a *P*-value passes a specific threshold additional weight. This could harm performance when this additional weight is unwarranted. For example, the correct answer to Question 1 is Option A regardless of the observed *P*-value presented; therefore, any additional weight given to *P*-values, 'statistical significance', or whether a *P*-value passes a specific threshold could harm performance.

Second, just as McShane and Gal (2016, 2017) asked academic researchers to explain why they chose the answer they chose in response to the question, we asked LLMs to do the same. As in McShane and Gal (2016, 2017), these explanations tended to emphasize specific thresholds and 'statistical significance'. For example, in one trial of Question 2 when the observed *P*-value was 0.049 and there was no prompt engineering, ChatGPT 4o responded, 'The *p*-value of 0.049 indicates that this difference is statistically significant at the conventional 5% significance level (since $0.049 < 0.05$), thereby providing evidence against the null hypothesis that there is no difference in recovery probability between the two drugs'. However, in one trial when the observed *P*-value was 0.051, it responded, 'The *p*-value of 0.051 indicates that there is not enough statistical evidence to reject the null hypothesis at a conventional significance level (e.g., 0.05)'. These responses suggest that dichotomania may be deeply ingrained in LLMs.

Although we cannot definitively ascertain that this is the case, we conducted a handful of trials for each model using the version of each question used in prior research but omitting the observed *P*-value

***Table 4.*** *o3 Mini results.*

| Prompt | Observed *P*-value | | | | |
|---|---|---|---|---|---|
| engineering | 0.01 | 0.049 | 0.051 | 0.099 | 0.101 |
| (a) Question 1: Description with response option preamble | | | | | |
| None | 100 | 100 | 96 | 32 | 15 |
| ASA | 100 | 100 | 99 | 78 | 61 |
| ASA P3 | 100 | 100 | 100 | 87 | 72 |
| (b) Question 1: Description without response option preamble | | | | | |
| None | 100 | 100 | 94 | 16 | 0 |
| ASA | 100 | 100 | 100 | 38 | 12 |
| ASA P3 | 100 | 100 | 100 | 87 | 34 |
| (c) Question 2: Prediction | | | | | |
| None | 100 | 100 | 0 | 0 | 0 |
| ASA | 100 | 100 | 1 | 1 | 0 |
| ASA P3 | 100 | 100 | 37 | 12 | 3 |
| (d) Question 3: Decision | | | | | |
| None | 100 | 100 | 18 | 1 | 0 |
| ASA | 100 | 100 | 23 | 1 | 0 |
| ASA P3 | 100 | 100 | 19 | 0 | 0 |

Note: Each cell of the table gives the number of times that the response was Option A for each question version and observed *P*-value across 100 independent trials. None denotes the version of the question used in prior research; ASA (ASA P3) denotes the version that employs prompt engineering with the six principles (Principle 3) of the ASA Statement.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

*p* (e.g., omitting 'A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a *p*-value of 0.01' from Question 2). The explanations nonetheless continued to emphasize 'statistical significance'. For example, in one trial of Question 2, ChatGPT 4o responded, 'The study shows a difference in recovery rates but it doesn't tell us if this difference is statistically significant'. In another trial, it responded, 'The study summary does not mention whether the difference is statistically significant'. The fact that the explanations continued to emphasize 'statistical significance' even when no *P*-value is presented further supports the notion that dichotomania may be deeply ingrained in LLMs and provides an opportunity for future research to more systematically examine when and in what manner such explanations make recourse to 'statistical significance'.

Third, so-called reasoning models such as OpenAI's o1 Mini, o1, and o3 Mini that automatically use a chain-of-thought approach in generating responses have recently been introduced (September 12, 2024, December 5, 2024, and January 31, 2025, respectively; OpenAI (2024a, 2024c, 2025a)). OpenAI notes that these models 'spend more time thinking before they respond (and) can reason through complex tasks' (OpenAI, 2024b). Although it is not entirely clear what that entails, key elements involve breaking down complex problems into component parts, generating multiple potential responses for each part, detecting and correcting mistakes within potential responses, choosing the best potential responses, and ultimately choosing the best overall response as the final one (Microsoft, 2025; OpenAI, 2024d; Woodie, 2025). OpenAI emphasizes that these models 'excel at' (OpenAI, 2024a) and have 'exceptional capabilities' (OpenAI, 2025a) in science, technology, engineering, and mathematics. Given this, we also assessed dichotomania in o3 Mini. Results can be found in Table 4. o3-mini performs dichotomously at the 0.05 threshold for Question 2 and Question 3 and worse on Question 1 than all versions of OpenAI's ChatGPT except ChatGPT 3.5 Turbo.

Consequently, it may be that new approaches to artificial intelligence that are radically different from LLMs may be required to improve performance (Sawers, 2025). Until then, it seems sensible for academic researchers to proceed with caution when using LLMs as a research tool.

**Acknowledgements.** We thank Will Thompson for exceptional research assistance.

**Supplementary material.** The supplementary material for this article can be found at https://doi.org/10.1017/jdm.2025.7.

## References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307.

Antu, S. A., Chen, H., & Richards, C. K. (2023). Using LLM (large language model) to improve efficiency in literature review for undergraduate research. In *LLM@ AIED*, pp. 8–16.

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, *16*(10), 335–338.

Cai, H., Cai, X., Chang, J., Li, S., Yao, L., Wang, C., Gao, Z., Wang, H., Li, Y., & Lin, M., et al. (2024). Sciassess: Benchmarking LLM proficiency in scientific literature analysis. arXiv preprint, arXiv:2403.01976.

Cash, T. N., Oppenheimer, D. M., & Christie, S. (2024). Quantifying uncertainty: Testing the accuracy of LLMs' confidence judgments. https://osf.io/preprints/psyarxiv/47df5_v1.

Fisher, R. A. (1935). Letter to the editor: Statistical tests. *Nature*, *136*, 474.

Freeman, P. R. (1993). The role of *p*-values in analysing trial results. *Statistics in Medicine*, *12*, 1443–1452.

Gans, J. (2025a). What will AI do to (p)research? https://joshuagans.substack.com/p/what-will-ai-do-to-presearch.

Gans, J. S.. (2025b). The efficient market hypothesis when time travel is possible. *Economics Letters*, *248*, 112209. https://doi.org/10.1016/j.econlet.2025.112209.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218.

Goodman, S. N. (2008). A dirty dozen: Twelve *p*-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140.

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology*, *186*(6), 639–645.

Greenland, S. (2019). Valid *p*-values behave exactly as they should: Some misleading criticisms of *p*-values and their resolution with *s*-values. *The American Statistician*, *73*(sup1), 106–114.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p* values, confidence intervals, and power: A guide to misinterpretations. *The American Statistician*, *70*(2), 1–12. (Online Supplement).

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1–20.

Holman, C. D. J., Arnold-Reed, D. E., de Klerk, N., McComb, C., & English, D. R. (2001). A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. *Epidemiology*, *12*(2), 246–255.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?*. National Bureau of Economic Research. Technical report.

Jansen, J. A., Manukyan, A., Al Khoury, N., & Akalin, A. (2025). Leveraging large language models for data analysis automation. *PloS one*, *20*(2), e0317084.

Jin, H., Zhang, Y., Meng, D., Wang, J., & Tan, J. (2024). A comprehensive survey on process-oriented automatic text summarization with exploration of LLM-based methods. arXiv preprint, arXiv:2403.02901.

Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT-4 from a human in a turing test. https://arxiv.org/abs/2405.08007.

Kapoor, S., Henderson, P., & Narayanan, A. (2024). Promises and pitfalls of artificial intelligence for legal applications. arXiv preprint, arXiv:2402.01656.

Kirkovska, A. (2024). LLM benchmarks: Overview, limits and model comparison. https://www.vellum.ai/blog/llm-benchmarks-overview-limits-and-model-comparison.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. https://arxiv.org/abs/2205.11916.

Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., & Kang, D. (2024). Benchmarking cognitive biases in large language models as evaluators. https://arxiv.org/abs/2309.17012.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., & Yin, Y., et al. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), AIoa2400196.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI scientist: Towards fully automated open-ended scientific discovery. arXiv preprint, arXiv:2408.06292.

Martin, L., Whitehouse, N., Yiu, S., Catterson, L., & Perera, R. (2024). Better call GPT, comparing large language models against lawyers. https://arxiv.org/abs/2401.16212.

McShane, B. B., & Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, *62*(6), 1707–1718.

McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, *112*(519), 885–895.

McShane, B. B., Bradlow, E. T., Lynch Jr. , J. G., & Meyer, R. J. (2024). "Statistical significance" and statistical reporting: Moving beyond binary. *Journal of Marketing*, *88*(3), 1–19.

Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., Zhang, M., Cao, C., Wang, J., & Wang, X., et al. (2024). The application of large language models in medicine: A scoping review. *Iscience*, *27*(5), 109713.

Microsoft. (2025). How reasoning models are transforming logical AI thinking. https://techcommunity.microsoft.com/blog/azuredevcommunityblog/how-reasoning-models-are-transforming-logical-ai-thinking/4373194.

Nechakhin, V., D'Souza, J., & Eger, S. Evaluating large language models for structured science summarization in the open research knowledge graph. *Information*, *15*(6), 328, 2024.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York, NY: Wiley.

OpenAI. (2024a). OpenAI o1-mini. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/.

OpenAI. (2024b). Introducing OpenAI o1-preview. https://openai.com/index/introducing-openai-o1-preview/.

OpenAI. (2024c). Introducing OpenAI o1. https://openai.com/o1/.

OpenAI. (2024d). Reasoning models. https://platform.openai.com/docs/guides/reasoning.

OpenAI. (2025a). OpenAI o3-mini. https://openai.com/index/openai-o3-mini/.

OpenAI. (2025b). Introducing deep research. https://openai.com/index/introducing-deep-research/.

Pearson, K. (Oct. 1906). Note on the significant or non-significant character of a sub-sample drawn from a sample. *Biometrika*, *5*(1/2), 181–183.

Sawers, P. (2025). Meta's Yann Lecun predicts 'new paradigm of ai architectures' within 5 years and 'decade of robotics'. https://techcrunch.com/2025/01/23/metas-yann-lecun-predicts-a-new-ai-architectures-paradigm-within-5-years-and-decade-of-robotics/?guccounter=1.

Tang, H., Zhang, C., Jin, M., Yu, Q., Wang, Z., Jin, X., Zhang, Y., & Du, M. (2025). Time series forecasting with LLMs: Understanding and enhancing model capabilities. *ACM SIGKDD Explorations Newsletter*, *26*(2), 109–118.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

https://doi.org/10.1017/jdm.2025.7 Published online by Cambridge University Press

Woodie, A. (2025). What are reasoning models and why you should care. https://www.bigdatawire.com/2025/02/04/what-are-reasoning-models-and-why-you-should-care/.

Xing, Y. (2024). Exploring the use of ChatGPT in learning and instructing statistics and data analytics. *Teaching Statistics*, *46*(2), 95–104.

Yu, Z., He, L., Wu, Z., Dai, X., & Chen, J. (2023). Towards better chain-of-thought prompting strategies: A survey. https://arxiv.org/abs/2310.04959.

Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., & Li, Q. (2024). Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge and Data Engineering*, *36*(11), 6889–6907. https://doi.org/10.1109/tkde.2024.3392335.

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, *634*(8032), 61–68.

**Cite this article:** McShane, B. B., Gal, D., and Duhachek, A. (2025). Artificial intelligence and dichotomania. *Judgment and Decision Making*, e23. https://doi.org/10.1017/jdm.2025.7