It must be very hard to publish null results

Ryan C. Briggs, Jonathan Mellon, Vincent Arel-Bundock*

2026-02-11

Publication practices in the social sciences act as a filter that favors statistically significant results over null findings. While the problem of selection on significance (SoS) is well-known in theory, it has been difficult to measure its scope empirically, and it has been challenging to determine how selection varies across contexts. In this article, we use large language models to extract granular and validated data on about 100,000 articles published in over 150 political science journals from 2010 to 2024. We show that fewer than 2% of articles that rely on statistical methods report null-only findings in their abstracts, while over 90% of papers highlight significant results. To put these findings in perspective, we develop and calibrate a simple model of publication bias. Across a range of plausible assumptions, we find that statistically significant results are estimated to be one to two orders of magnitude more likely to enter the published record than null results. Leveraging metadata extracted from individual articles, we show that the pattern of strong SoS holds across subfields, journals, methods, and time periods. However, a few factors such as pre-registration and randomized experiments correlate with greater acceptance of null results. We conclude by discussing implications for the field and the potential of our new dataset for investigating other questions about political science.

*We thank Eve Fournier, Florence Laflamme, Jayden Lakhani-Travis, Truong Pham, Emma Pirard, Andrews Sai, Tim Larson, and Zhixiang Wang for research assistance. We thank Abel Brodeur, Devon Cantwell-Chavez, Rex Douglass, James Fahey, Bob Kubinec, Carlisle Rainey, and participants at SPSA and APSA for advice and discussion. We acknowledge funding from the Social Sciences and Humanities Research Council of Canada (grant awarded June 2023). We acknowledge the use of LLMs as brainstorming, editing, information retrieval, and coding aids in the production of this manuscript.

Most of the statistical tests that political scientists conduct are severely underpowered (Arel-Bundock et al. 2026). That simple observation implies that the majority of null hypothesis significance tests we run should yield null results. Yet, the published literature tells a very different story. When reading the pages of peer-reviewed journals in political science, one must wonder: How can we (nearly) always be right? Why does the data so consistently support our hunches? Is it so hard to publish null results?

In this article, we document a massive gap between the *expected* and the *observed* rates of null results in published political science research. We argue that the central mechanism behind this gap is selection on significance (SoS), that is, a filter that makes statistically significant estimates more likely to appear in print than null results.

SoS can operate through many mechanisms. Researchers could quickly abandon tests that produce null results, or engage in specification searching and p-hacking, or otherwise get lost in the "garden of forking paths" (Gelman and Loken 2013). Authors could strategically choose to highlight significant results but relegate null results to a footnote or appendix. Editors and reviewers may have a taste for "positive" findings and so be more likely to reject papers with null results. Or authors may believe that Editors or reviewers hold such a preference, and then only submit papers they expect will be accepted.

SoS has profound consequences. When significant results are more likely to make it to print, the published literature ceases to be representative of the evidence generated: published estimates are afflicted by a "winner's curse" that biases them away from zero, researchers commit more Type I error, meta-analyses become unreliable, and published studies are less likely to be replicable. When null results remain hidden in the file drawer, researchers may waste time and money studying the same ineffective interventions over and over (Rosenthal 1979; Van Zwet and Cator 2021; Gelman and Carlin 2014; Gelman and Tuerlinckx 2000; Gerber and Malhotra 2008; Berinsky, Druckman, and Yamamoto 2021; Vasishth and Gelman 2017; Irsova et al. 2023).

Concerns over SoS have been voiced for decades (Sterling 1959) but, as we discuss in Section 1, the empirical literature in political science is still limited in important ways. Indeed, the best studies to date focus on small areas of the discipline where data are especially rich. These studies are important, but their limited scope means that we do not yet have a clear sense of the extent of the problem in the discipline at large.

In this article, we present the first large-scale, discipline-wide evidence describing the extent of SoS in political science and the factors associated with it. We use large language models (LLMs), backed by extensive human validation, to extract structured information from the full texts of 101,475 articles published since 2010 in 156 peer-reviewed political science journals.

Using this newly constructed dataset, we measure the share of null, non-null, and precise-null claims that authors choose to highlight in their articles' abstracts. We find that the share of pure null results is extremely small: fewer than 2% of abstracts report only null findings. By contrast, over 90% of articles that rely on statistical methods prominently claim to reject at least one null hypothesis.

To benchmark these observations, we develop a simple model of publication bias and conduct a calibration exercise. This framework highlights a large gap between the number of null results that statistical theory would lead us to expect and the number that appear in the published record. With generous assumptions about statistical power and the prevalence of true effects, we find that estimates that are statistically distinguishable from zero must be at least one order of magnitude more likely to appear in print than null results. Arguably more realistic assumptions can easily produce magnitudes of selection on significance around 100x. These results suggest that political science has a high degree of SoS.

We then leverage granular metadata extracted by LLMs to examine how these patterns vary across contexts. Several contextual elements are plausibly at play here. For example, differences across journals may occur due to varying evidentiary standards or acceptance rates; temporal variation could emerge because of shifting norms and open science reforms; methodological differences could speak to statistical power. These ideas are all reasonable *a priori* but, in fact, we show that a strong selection filter operates across subfields, journals, methods, and time periods. We do find that pre-registered studies and experiments are somewhat more likely to report null or mixed findings, but these differences are modest in magnitude and are dwarfed by the scale of SoS.

This article makes five distinct contributions. We provide the first discipline-wide, article-level measurement of selection on significance in political science. We do this at a scale that far exceeds prior empirical efforts, covering fifteen years, more than 150 journals, and nearly the entire universe of recent peer-reviewed research in the field. Second, we introduce a transparent accounting framework that links statistical power to the observed prevalence of null results, yielding interpretable estimates of selection on significance. Third, we show that the gap between expected and observed null results is remarkably stable across journals, subfields, methods, and time, suggesting that SoS is a structural feature of the discipline rather than a localized problem. Fourth, we demonstrate that carefully validated large language models can match or exceed highly trained human coders in this setting when extracting nuanced, theory-laden information from complex scholarly articles, enabling meta-scientific analysis at unprecedented scale. Finally, we publish a free, richly annotated dataset that codes a wide range of article characteristics, including methods, subfields, open science practices, and geographic and temporal coverage. These data open the door to many future research projects, well beyond the study of publication bias.

To be clear, the publication filter that we document here is not a benign pathology. It is a structural feature of our reporting practices that threatens the credibility and accumulation of knowledge in political science.

1 SELECTION ON SIGNIFICANCE

We say that there is selection on significance when the publication of a scientific finding depends on whether its associated test statistic crosses a conventional threshold, such as $|t| > 1.96$. SoS can operate between or within studies. At the article level, editors, reviewers, and authors could prefer to publish papers that report statistically significant findings. Within papers, authors could choose to emphasize statistically significant findings and downplay hypothesis tests that fail to reject the null. For example, they could highlight significant results in the abstract, but relegate tests with large p values to a footnote or appendix.

As we noted in the introduction, SoS is problematic because it distorts the scientific record and wastes resources. SoS also undermines statistical inference in at least three ways: biasing published estimates away from zero (Van Zwet and Cator 2021; Gelman and Carlin 2014), reshaping the distribution of reported signs and magnitudes and thus promoting overconfidence in replicability (Vasishth and Gelman 2017; Brodeur et al. 2016), and undermining meta-analytic credibility (Irsova et al. 2023; J. P. A. Ioannidis 2005). To gauge the severity of these problems, we must necessarily understand the scope and magnitude of SoS in actual practice.

Prior work on SoS can be grouped into four broad categories. First, some authors do not adopt an empirical lens like ours, but instead rely on simulations or formal models to characterize the severity of SoS and its mechanisms. Simmons, Nelson, and Simonsohn (2011) use computer simulations to illustrate how easy it can be for researchers to generate statistically significant false positives by manipulating seemingly innocuous features of the research design. Formal theorists have also modeled SoS using different behavioral assumptions for authors, reviewers, and publishers (Hedges 1992; Andrews and Kasy 2019; Berinsky, Druckman, and Yamamoto 2021). A key insight of this literature is that, under selection, only limited features of the data-generating process are learnable from the observed record alone. This insight informs our approach: in the rest of this paper, we will *not* attempt to disentangle the mechanisms that generate SoS, and we will not produce causal estimates of the effect of statistical significance on the probability of publication. Instead, the results reported in Section 4 should be interpreted as capturing a host of behavioural effects on the part of authors as well as editorial or reviewer selection against null results.

Second, some of the strongest direct evidence comes from studies with access to exceptionally rich datasets in narrow domains, often including confidential data on both executed and submitted tests. Franco, Malhotra, and Simonovits (2014) study Time-sharing Experiments for the Social Sciences (TESS) and show that null results are less likely to be eventually published, in large part because they are less likely to be written up and submitted in the first place. Franco, Malhotra, and Simonovits (2015) uses data from the same program to show that authors routinely fail to report all the experimental conditions in their research design. Moniz, Druckman, and Freese (2025) find similar patterns, though with some evidence of improvements over time. With a narrower focus but even richer data, Brodeur et al. (2023) uses confidential information on journal submissions to show that while initial author submissions exhibit significant bunching consistent with p-hacking, peer review itself has little effect on test statistic distributions. These studies provide the most direct evidence of SoS, but their scope is limited, often confined to a single journal, repository, or intervention type. This makes it difficult to assess whether their findings generalize to the broader discipline.

Third, several scholars have looked at the distribution of published p values and effect sizes to infer the presence of SoS. Brodeur et al. (2016) analyze the distribution of p values published by three high-profile economics journals and find that this distribution is camel-shaped, which strongly suggests the presence of SoS. Gerber and Malhotra (2008) present similar results for political science. A related strand of research starts from meta-analytic estimates to estimate the inflationary bias induced by SoS in specific literatures (J. Ioannidis, Stanley, and Doucouliagos 2017; Arel-Bundock et al. 2026). While these studies tend to cover more articles than those relying on confidential data, they are still limited in scope, typically focusing on a small number of “top” journals or specific literatures, and they do not typically assess how SoS could vary in different contexts.

Fourth, some studies have investigated whether the degree of SoS varies systematically across contexts. The “publication advantage” of small p values may plausibly differ based on subject areas, time periods, methodology, open science practices, and publication venues. Some evidence suggests methodological heterogeneity: Brodeur, Cook, and Heyes (2020) analyze 25 economics journals and find that SoS varies by empirical method but not by journal prestige; Vivalt (2019) analyzes a database of impact evaluations and finds similar patterns for articles using observational causal inference methods but less SoS for randomized experiments. However, existing evidence does not conclusively establish these patterns across the broader discipline. Another key question is whether open science reforms have reduced SoS. Policies to reduce SoS vary in their stringency (Ofosu and Posner 2023), and available evidence suggests more improvement when reforms more stringently target SoS at all stages of the research pipeline. For example, Brodeur et al. (2024) find little evidence of pre-registration working without a clear pre-analysis plan. The most stringent policy to reduce SoS is the registered report (Chambers and Tzavella 2022). While these are new, an early stock-taking exercise in psychology found that while standard psychology articles had 96% significant results, in registered reports this fell to 44% (Scheel, Schijen, and Lakens 2021). However, the literature to date has not systematically examined heterogeneity in SoS across the full range of contextual factors that might matter, has not done so in political science, and has not examined these patterns across the full scope of an entire discipline.

To study selection on significance across an entire discipline and measure how it varies across contexts, we need detailed metadata on each article. This includes information on subfield, methodology, publication venue, open science practices, and whether papers contain null or non-null hypothesis tests. Existing studies lacked access to such detailed metadata at scale. We use large language models to conduct metascience at this scale, extracting structured information from the full text of nearly 100,000 articles. Thus, our paper contributes to the existing literature on publication bias by assembling a large-scale dataset that combines rich article-level metadata with substantially broader coverage of journals and articles than has previously been available.

2 LARGE LANGUAGE MODELS FOR INFORMATION RETRIEVAL

A major barrier to studying selection on significance at scale is data acquisition: extracting structured information from academic articles is labor-intensive and expensive. Unfortunately, workers recruited on platforms like Mechanical Turk often lack the substantive and technical expertise to answer detailed questions about frontier research in the social sciences. Successful human-led coding efforts thus tend to require domain experts at the graduate student level or above. Some meta-science studies have adopted this strategy, employing large teams of expert research assistants to read many articles, but the cost of such efforts is extremely high (Brodeur, Cook, and Heyes 2020).

We address this challenge using large language models. Since the introduction of ChatGPT in late 2022, academics have rushed to take advantage of the abilities of modern LLMs. These models appear proficient at many common research data tasks, including extraction of information from text at scale (Mellon et al. 2024; Huang, Kwak, and An 2023; Gilardi, Alizadeh, and Kubli 2023). These tools can handle quite complicated instructions and contexts, and so can also be part of complex information treatments (Argyle et al. 2023; Velez 2024). However, this rush to adopt LLMs has also been accompanied by warnings that LLMs may be less effective than existing techniques (Bisbee et al. 2024; Heyde, Haensch, and Wenz 2024), as well as concerns about lock-in to proprietary providers (Spirling 2023) and replicability (Barrie, Palmer, and Spirling 2024).

Our study complements the LLM work of other teams. For example, Grossman et al. (2024) and Grossman, Dinneen, and Torreblanca (2025) in political science and Garg and Fetzer (2025) in economics use LLMs to extract structured information from many academic articles. LLMs are also being tested for other meta-science tasks such as replication (Brodeur et al. 2025).

Three features distinguish our approach from past work. First, we benchmark both human research assistants and LLMs against expert-coded ground truth that was produced through a reconciliation exercise where disagreements were adjudicated by faculty experts. This matters because LLMs can outperform human Research Assistants (RA) in some tasks, so disagreement between RAs and LLMs does not, by itself, imply that the LLM is wrong (Yang et al. 2025). Second, our pre-registered validation presents measures of sensitivity and specificity rather than only accuracy. This is important because raw accuracy can be misleading when coding rare fields. Third, we report raw results and results from a pre-registered Bayesian model. This helps reduce noise in small samples and so increases the power of our tests of LLM performance.

3 MEASUREMENT AND VALIDATION

In this section, we describe our data sources, what we measure, and how we coded variables using both research assistants and LLMs. We also describe our extensive data validation process and suggest that, in this information retrieval task, LLMs can outperform trained research assistants at a fraction of the cost.

3.1 Data

Our target population includes well-recognized peer-reviewed political science journals from 2010 forward.[1] To identify them, we merged high-impact journal lists from Clarivate's Journal Citation Reports and Google Scholar Metrics across Political Science, International Relations, and Public Administration. After removing non-peer-reviewed outlets and journals flagged for questionable practices, we were left with a final list of 205 journals.[2] We do not scrape publisher websites for full article texts. Instead, we obtained them through formal data sharing agreements with publishers, who sent us high-quality XML/HTML that can be parsed reliably. At present, the dataset includes 101,475 articles published since 2010 from 156 of the 205 journals in the sampling universe.

[1] This start date is necessarily arbitrary, but qualitative reviews of the literature suggest that reporting conventions and file formats were still unsettled before 2010, making reliable data extraction more difficult for earlier periods.

[2] See Appendix B for detailed selection criteria.

Table 1: Information extracted from each article

Category	Examples
In sample	Original research, book review, editorial
Results	Null, non-null, mixed, reject meaningful effect
Subfield	Comparative, American, IR, theory, methodology
Methodology	Survey experiment, lab experiment, field experiment, interview, regression, case study, ethnography, instrumental variables, formal theory, etc.
Open science	Data availability, code availability, pre-registration
Coverage	Countries studied, start year, end year

3.2 What we measure

We use a codebook to extract a wide range of information from each article, as summarized in Table 1.

Our primary goal is to characterize whether published articles present their central findings as null, non-null, or mixed. A typical paper contains many hypothesis tests serving different purposes: some represent the main claims, others assess robustness, explore additional patterns, or act as placebo tests where null results are expected and even supportive of the argument. In that setting, mechanically extracting z-statistics or p values does not capture how authors frame the overall contribution. We therefore focus on article abstracts, the primary communicative venue where authors summarize their main findings for readers, editors, and reviewers.

Our precise coding rules for classifying abstracts as reporting null, non-null, or mixed results are presented in Appendix C. Here, we outline the key strengths and limitations of this approach. The main advantage is that it defers to authors' own judgments about which results matter most. If authors believe that null results are harder to publish and therefore strategically emphasize or introduce non-null findings to improve their chances of acceptance, that behavior is itself part of the selection process we aim to measure. Abstracts also have practical advantages: they are short, standardized, and easy to identify, making them well suited for coding by both research assistants and large language models.

The main limitation is that authors may omit null results from abstracts even when those results appear in the paper. This represents a form of selection on significance that operates within papers rather than across papers. Although this is likely less severe than the complete non-submission of null-result papers, it can still distort the scientific record, especially if readers rely on abstracts to form initial beliefs.[3] Thus, it is important to emphasize that our estimates reflect the combined influence of author behavior (such as file-drawer decisions or repackaging results) and journal behavior (publication bias). These mechanisms cannot be cleanly separated in our data, and we therefore interpret our results as capturing their joint effect.

[3] Researchers very often only interact with abstracts (Nicholas, Huntington, and Jamali 2007) and among all academics this "scanning" behaviour seems to be most common among social scientists (Palmer, Teffeau, and Pirmann 2009).

Beyond null and non-null results, our codebook captures rich metadata about each article, which enables us to examine how publication patterns vary across important dimensions of the discipline. As shown in Table 1, we extract information on subfield (comparative politics, American politics, international relations, political theory, methodology), methodology (surveys, experiments, regression, case studies, ethnography, formal theory, etc.), open science practices (data and code availability, pre-registration, etc.), and geographic and temporal coverage (countries studied, start and end years). These fields are particularly valuable because they allow us to test whether the strong patterns of selection on significance we observe are consistent across different methodological approaches, subfields, and institutional contexts. The heterogeneity (or lack thereof) in null result reporting across these categories provides insight into the mechanisms driving publication bias and helps identify contexts where scientific communication is more or less distorted. Moreover, this structured dataset extends far beyond our investigation of publication bias, offering researchers a comprehensive resource for studying methodological trends, geographic coverage patterns, and the diffusion of open science practices. Section 7 highlights this potential by discussing several fruitful avenues for future research.

3.3 Codebooks for humans and for machines

A common codebook was used to guide both the RA and LLM coders in information extraction, with some additional interpretation notes shared with RAs. Designing the codebook was an iterative process: we started with a draft, had RAs and LLMs code a small number of articles, evaluated the output, revised the codebook, and repeated this cycle many times over a period of months. Four human RAs, all graduate students in social science programs, accumulated months of practice during this development phase before coding the actual validation sample. Together RAs coded over 2,500 person-articles in the pilot phase, initiated 35 distinct email chains about unclear feature descriptions, and received regular feedback from a PI.

3.4 Validation

To benchmark RA and LLM performance, we drew a stratified random sample with one article per journal for each 5-year period between 2010 to 2024.[4] Each article in this validation set was first coded by two humans and the state-of-the-art model of the time: OpenAI's GPT-5.[5] If all three coders agreed on a field, we treated their unanimous entry as the "ground truth." In case of disagreement, one of the principal investigators reconciled conflicting entries, often after extensive team discussion. This reconciliation was done blindly, without knowing which values were proposed by the LLM or RAs. After establishing the ground truth in the validation set, we re-coded all the sampled articles using two more

[4] More details on the validation process are reported in Appendix B. The validation analysis follows a pre-registered confirmatory analysis plan, with minor deviations documented in Section A: https://osf.io/fra5d

[5] On the day we conducted the comparative performance evaluation (2025-12-18), Kimi 2 Thinking ranked first in the open weights category on the benchmarks run by Artificial Analysis (2025) and second on the ranking published by White et al. (2025).

LLMs: the best available open weights model at the time (Kimi K2) and a cheaper version of the OpenAI model (GPT-5 mini).

To measure the performance of RA and LLM coders, we begin by computing standard metrics like accuracy, precision, sensitivity, and specificity. These statistics, reported in Appendix B, show that LLMs typically outperform human RAs on most metrics, for nearly all of the fields we measure. The strong relative performance of LLMs is especially noticeable for sensitivity. Anecdotally, we found that RA performance on this data extraction task often suffers because doing it well requires sustained focus to search for many rare events. Moreover, the RAs sometimes expressed frustration because they found the task boring, which made it hard for them to maintain focus for hours on end.

The descriptive results in Appendix B show that our open weights model performed considerably worse than both proprietary models. However, GPT-5 mini performed nearly as well as GPT-5, despite being much cheaper to use. Based on this observation, we decided to deploy GPT-5 mini as our model for coding the full sample.

To increase the statistical power of our validation, to represent uncertainty, and to provide an interpretable global summary of performance, we then fit a Bayesian hierarchical regression model that pools information across coders and variables.[6] Figure 1 reports the posterior distribution of sensitivity and specificity for human coders and GPT-5 mini, pooled across all articles and fields.

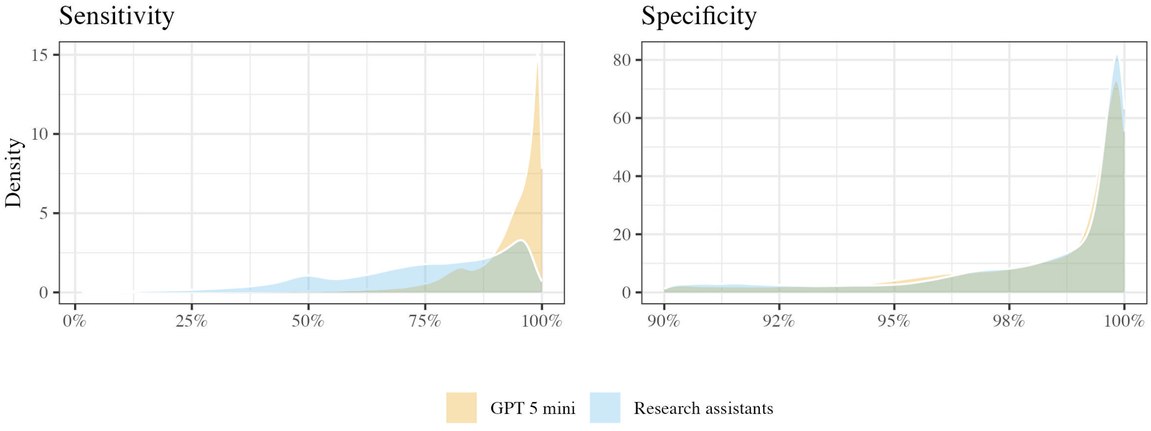

Figure 1: Distribution of sensitivity and specificity scores for human coders and GPT-5 mini

This plot makes clear that both humans and LLMs are very accurate at identifying the absence of features (specificity). While this is good news, it is not especially impressive. Indeed, many of the fields that we asked coders to extract are rare, so a naive classifier that would say "no" everywhere would score well on this metric. What is more impressive is that our coders generally perform very well on the sensitivity metric. Moreover, Figure 1 shows that GPT-5 mini displays considerably higher sensitivity than the RAs,

[6] Appendix B describes this model in detail.

who have a long left tail of low sensitivity features. The difference in sensitivity between the LLMs and pooled RAs (with 95% credible interval) is -17.10 [-35.04, -1.91].

In sum, through an extensive (and costly) validation process, we have demonstrated that GPT-5 mini performs very well at recovering the ground truth data. It is clearly better than highly trained graduate students at this specific information retrieval task. Moreover, we estimate that using this LLM to code a single paper costs roughly 1000 times less than an RA, and we note that LLMs can process papers much faster by making parallel API calls.

4 HYPOTHESIS TESTS ARE COMMON, BUT NULL RESULTS ARE RARE

Having validated the effectiveness of LLM coding of information contained in political science publications, we now turn to our primary goal: measuring the prevalence of null and non-null results across the discipline. We leverage a large sample of 101,475 articles from 156 journals, which allows us to systematically examine how selection on significance operates at scale.

Figure 2 shows how the use of statistics varies over time and across journal rank.[7] In the top three political science journals, the share of all in-scope articles that use statistical inference has held steady at about 80%. In all other journals, the share has risen considerably over time, from around 30% in 2010 to just under half of all articles by 2024. Political science has become an increasingly statistical discipline, which makes it more important than ever for our field to ensure that the published record faithfully represents the statistical tests that researchers actually conduct.

Unfortunately, we find that the published record tells a strikingly one-sided story. 94% of abstracts explicitly report a non-null result, while only 18% report a null result.[8] In very few articles (2%) do authors report only a null result in the abstract. Only 0.42% of abstracts claim to reject meaningful effects (e.g., via an equivalence test or by conducting a well-powered study and reporting a null result with a tight confidence interval).[9] In sum, almost every published statistical article in the discipline prominently claims to reject a null hypothesis.

5 NULL RESULTS SHOULD BE MUCH MORE COMMON

Is it odd that political science journals report so many statistically significant results? Yes, it is! But how odd is it, exactly?

We answer this question by developing and calibrating a simple model of publication bias. This allows us to estimate how many null results we should expect to see, and to quantify the gap between our theoretical expectations and the observed data. This exercise suggests that SoS must be very strong.

[7] The codebook is in Appendix C. In brief, we define an article as statistical if it ever quantifies uncertainty in an empirical estimate.

[8] These figures exceed 100% because some articles report both null and non-null results.

[9] The statistics reported in this paragraph refer to in-scope statistical articles across all journals.

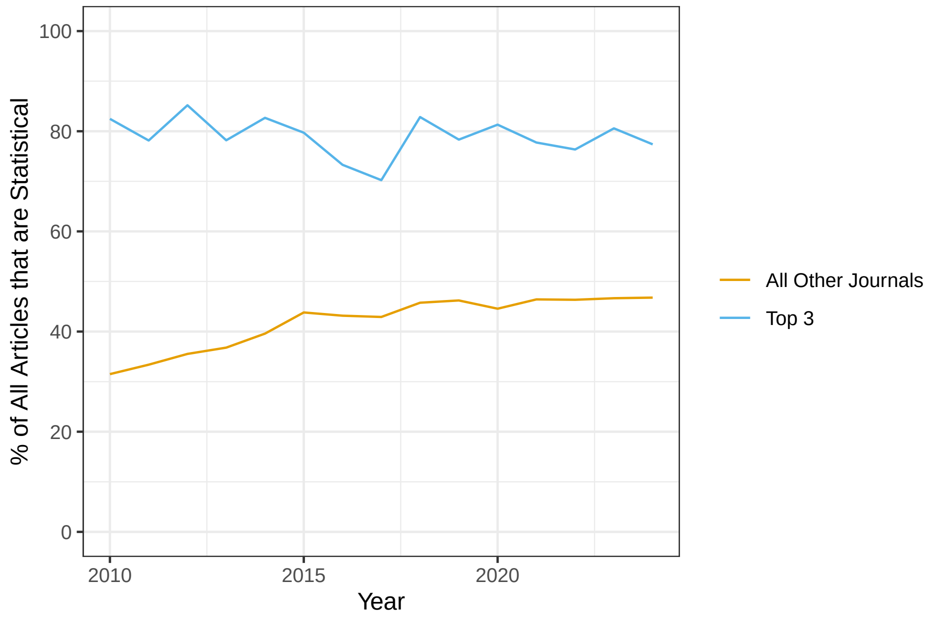

Figure 2: Share of articles that include a statistical analysis.

5.1 *Definitions*

To characterize the severity of SoS, we start by distinguishing five types of abstracts: (e) *empty* abstracts that make no statistical claim, either null or significant; (n) *null-only* abstracts that mention only null results; (s) *significant-only* abstracts that mention only significant results; (m) *mixed* abstracts that mention at least one significant and at least one null result; and (u) abstracts that remain *unpublished*.

We use τ_j with $j \in \{e, n, m, s\}$ to denote the shares of each abstract type in our published sample: empty ($\tau_e \approx 0.05$), null-only ($\tau_n \approx 0.02$), mixed ($\tau_m \approx 0.16$), significant-only ($\tau_s \approx 0.77$).

ρ stands for the probability that any given test yields a null result. This probability is determined by statistical power (ϕ), the share of tests that target true non-zero effects (π), and the chosen rate of Type I error (α). Later on, we will show how results change when these parameters vary. But, for now, it is useful to tie them to specific numeric values.

For statistical power, we use the average discipline-wide estimate of $\phi = 0.25$ from Arel-Bundock et al. (2026). Since it is impossible to pin down π empirically, we make the generous assumption that three quarters of the hypotheses that political scientists choose to investigate are correct ($\pi = 0.75$). We fix the significance level to a conventional threshold ($\alpha = 0.05$).

These (reasonable) numbers imply that the share of null results in political science should be high:

$$\begin{aligned}\rho &= 1 - \pi\phi - (1-\pi)\alpha \\ &= 1 - (0.75)(0.25) - (1-0.75)(0.05) \\ &= 0.80\end{aligned}$$

Let $S_i \in \{0,1\}$ indicate whether test result i is statistically significant. Imagine that an investigator conducts k independent statistical tests.[10] $M = \sum_{i=1}^{k} S_i$ is the number of significant results, with $M \sim \text{Binomial}(k, 1-\rho)$. The probability that all k tests yield null results is $\text{Pr}(M = 0) = \rho^k$.

Finally, let $R_i \in \{0,1\}$ record if test result i is reported in the abstract of a published article.

5.2 Selection

We now introduce the two most important parameters in our framework: a and b. These parameters characterize the SoS filter under two regimes: when some results are significant, and when all results are null.

The probabilities of reporting a null result in those two regimes are:

$$\begin{aligned}a &= \text{Pr}(R_i = 1 \mid S_i = 0,\ M \geq 1) \\ b &= \text{Pr}(R_i = 1 \mid S_i = 0,\ M = 0)\end{aligned}$$

Distinguishing these two probabilities allows us to capture a key intuition about scientific reporting. When a study has at least one significant result ($M \geq 1$), there is a salient "positive" finding to anchor the author's narrative. In that context, the relevant behavioral margin is whether null results are also reported, alongside the significant result. In contrast, when a study has no significant results ($M = 0$), there is no "positive" finding to anchor reporting. In that case, the relevant behavioral margin becomes whether any test result is reported at all.

This split between two regimes is intuitive, and it serves a useful role in identification: since the data we observe is binned into "only significant," "mixed," and "only null", there are two independent moments. Using a single inclusion probability would force the same behavior in $M \geq 1$ and $M = 0$, mechanically tying the frequency of mixed reports to the frequency of null-only reports. Such a model would be far too rigid.

[10] In practice, tests within a paper are often correlated rather than independent. Correlation reduces the effective number of independent tests, $k_{\text{eff}} \leq k$. At one extreme, with no correlation, $k_{\text{eff}} = k$. At the other extreme, with perfect correlation, all tests are essentially the same test, so $k_{\text{eff}} = 1$—which reduces to the single-test accounting framework above. The results we present here thus span both limiting cases: $k = 1$ corresponds to perfectly correlated tests, and higher values of k correspond to increasingly independent tests.

5.3 *Calibrating b: all results are null ($M = 0$)*

We now introduce notation to distinguish empirical shares from unconditional masses. As before, τ_j denotes the published-sample shares in our data. In contrast, U_j denotes the unconditional model-implied mass over all studies, including the unpublished or unobservable ones. These objects are linked by $\tau_j = U_j/(1 - U_u)$.

When $M = 0$, all tests yield null results. In that regime, the unconditional probability of each abstract category depends on b and k. Recall that ρ^k is the probability that all k independent tests yield null results, and $(1 - b)^k$ the probability that k null results go unreported. Then,[11]

$$\begin{aligned} U_u + U_e &= \rho^k (1-b)^k && \text{Unpublished or empty} \\ U_n &= \rho^k \left[1 - (1-b)^k\right] && \text{Null only} \end{aligned}$$

We can tie these unconditional masses to the observed data with:

$$\frac{\tau_n}{1 - \tau_e} = \frac{U_n}{1 - U_u - U_e} = \frac{\rho^k \left[1 - (1-b)^k\right]}{1 - \rho^k (1-b)^k}.$$

Solving for b:

$$b = 1 - \left(\frac{\rho^{-k} \left(\rho^k \tau_e - \rho^k + \tau_n\right)}{\tau_e + \tau_n - 1} \right)^{1/k} \tag{1}$$

Plugging in the observed shares of abstracts for τ_n and τ_e, the number of tests $k = 5$, and $\rho = 0.8$ gives us this numeric estimate:

$$b \approx 0.008.$$

When a study finds no significant result across its 5 tests, any given null result has less than a 1% chance of being reported. The probability that the abstract of this study is published and mentions any null result at all is $1 - (1 - b)^k \approx 3.8\%$.

[11] In this regime, significant-only and mixed outcomes are zero by construction: $U_s = U_m = 0$.

5.4 *Calibrating a: some results are significant ($M \geq 1$)*

When $M \geq 1$, some tests yield statistically significant results. For simplicity, we normalize the probability of reporting a significant result to 1: if $S_i = 1$, then $R_i = 1$.[12] By this normalization, unobserved and null-only outcomes are impossible when $M \geq 1$. Hence, we focus on:

$$\begin{aligned} U_s(a) &= \sum_{m=1}^{k} \Pr(M = m)\,(1-a)^{k-m} && \text{Significant only} \\ &= \sum_{m=1}^{k} \binom{k}{m}(1-\rho)^m \rho^{k-m}(1-a)^{k-m} \end{aligned}$$

As above, match the unconditional probabilities to observed shares of abstracts:

$$\frac{\tau_s}{1-\tau_e} = \frac{U_s(a)}{1-U_u-U_e} = \frac{\sum_{m=1}^{k}\binom{k}{m}(1-\rho)^m\rho^{k-m}(1-a)^{k-m}}{1-\rho^k(1-b)^k}. \tag{2}$$

Given the value of b in Equation 1, $\rho = 0.8$, $k = 5$, and the observed values of τ_s and τ_e, we can use numerical methods to solve for

$$a \approx 0.05.$$

Conditional on a study having at least one statistically significant result, any given null result has only about a 5% chance of being reported in the abstract. Paper-level implications follow directly.[13] With $k = 5$ and $a \approx 0.05$, a study with 1 significant result and 4 null findings has about a 19.5% chance of presenting a mixed abstract.

[12] This normalization assumes that whenever a study contains at least one statistically significant result, the abstract will mention at least one. Relaxing this assumption typically makes the implied level of SoS worse. To see how, consider the critical violation: a study that finds a significant result, fails to report it in the abstract, but reports some null result(s) instead. If such cases are common, they shift observations from the $M \geq 1$ regime onto the null-only mass. To match the observed null-only share, the calibrated value of b must be even lower.

[13] To compute paper-level estimates, we assume that R_i are independent across tests, conditional on the selection regime ($M = 0$ vs. $M \geq 1$). This assumption would be violated if reporting some tests crowds out others, or if authors "bundle" test reporting.

5.5 *Sensitivity to assumed parameters*

In the previous sections, we computed a and b using specific numeric assumptions. Now, we show that our main conclusions remain broadly unchanged under reasonable alternatives. The key parameters that govern a (Equation 2) and b (Equation 1) are k and ρ.

The number of tests k is important, because it determines the probability of obtaining at least one significant result, which in turn determines the relative importance of the two selection regimes. We will consider what happens to the SoS parameters a and b in hypothetical studies that include between 1 and 10 "main" tests that are candidates for inclusion in the abstract.

The probability of obtaining a null result, ρ, is crucial because it dictates the relative frequency of null and significant results. This, in turn determines the relative importance of the two selection margins. In political science, it seems likely that ρ is high, because statistical power is generally poor (Arel-Bundock et al. 2026). Figure 13 of the appendix shows how ρ varies with ϕ and π. For example, if statistical power is low ($\phi = 0.1$) and political scientists have tremendous theoretical intuition ($\pi = 0.9$), the expected share of null results should be roughly 90%. If statistical power is implausibly high ($\phi = 0.5$) but our theories are coin flips ($\pi = 0.5$), then we should get null results roughly 70% of the time.

The top panel of Figure 3 shows that as soon as k becomes moderately large, authors become very unlikely to report null results alongside significant results in their abstracts. Instead, the SoS filter inflates the number of significant-only reporting, at the expense of mixed results.

The bottom panel shows that even when k is high and ρ is low, the probability that an author will publish null-only abstracts remains very low (<10%). More optimistic assumptions about power (ϕ) and the prevalence of true effects (π) generally reduce ρ and reduce the implied severity of SoS, but SoS remains extremely strong under all reasonable parameter combinations.

Our substantive conclusion is thus robust: At some point in the process from original testing to publication, null results are being extensively filtered out. It must be—or authors must perceive it to be—very hard to publish null results.

6 CONTEXTUAL FACTORS DO NOT MATTER VERY MUCH

The evidence presented so far establishes that selection on significance is remarkably strong in political science. Having demonstrated this pervasive pattern, we now dig deeper to assess whether this bias varies systematically across different subsets of the data. Our goal is to identify contexts where the scientific record may be more or less distorted by SoS, and to provide clues for further investigation into the mechanisms that drive publication bias.

There are reasons to expect variation. For example, past research found that p-hacking was less extreme in randomized experiments (Vivalt 2019; Brodeur, Cook, and Heyes 2020) so perhaps SoS is lower there. There could also be variation across subfields as research practices, data availability, and methods vary by subfield. Journal prestige could be a factor, either with top journals having more stringent review

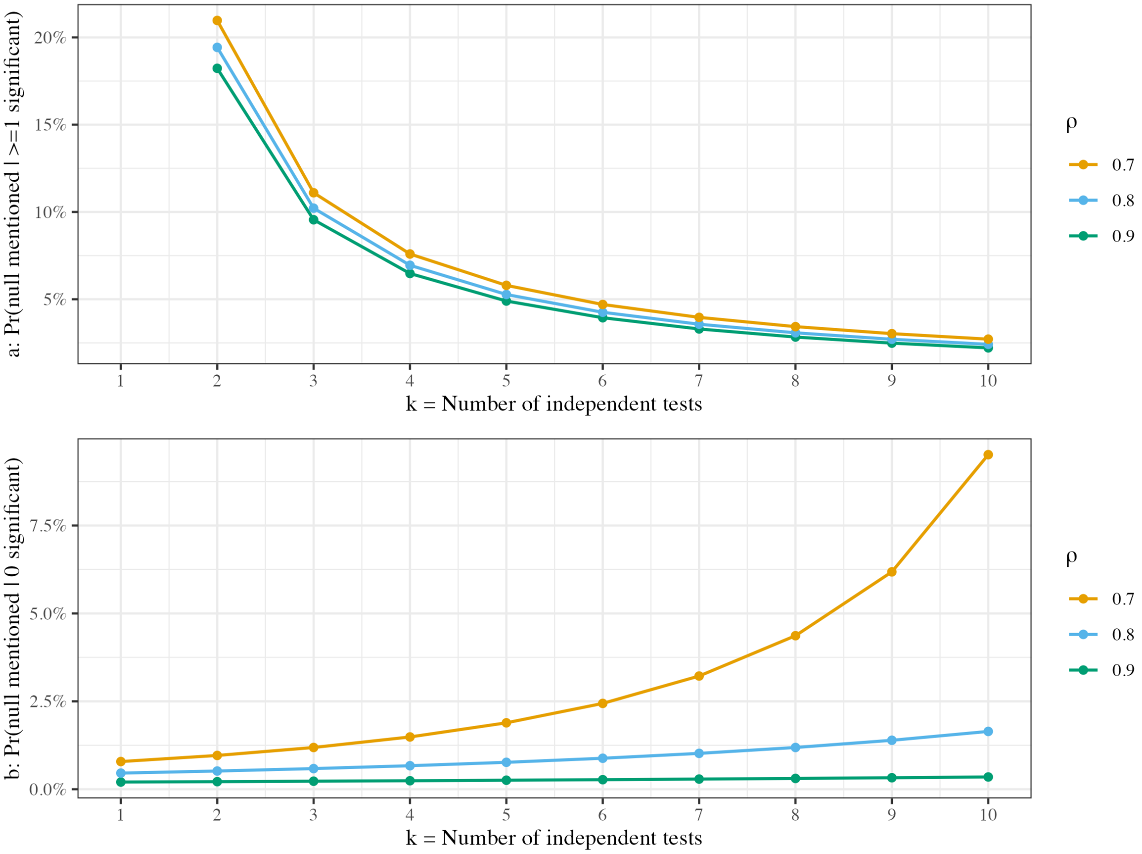

Figure 3: Calibrated reporting probabilities a and b over the number of tests k, with lines colored by ρ.

or succumbing to a dynamic where the demand for ideal papers ends up with stronger selection for significant results.

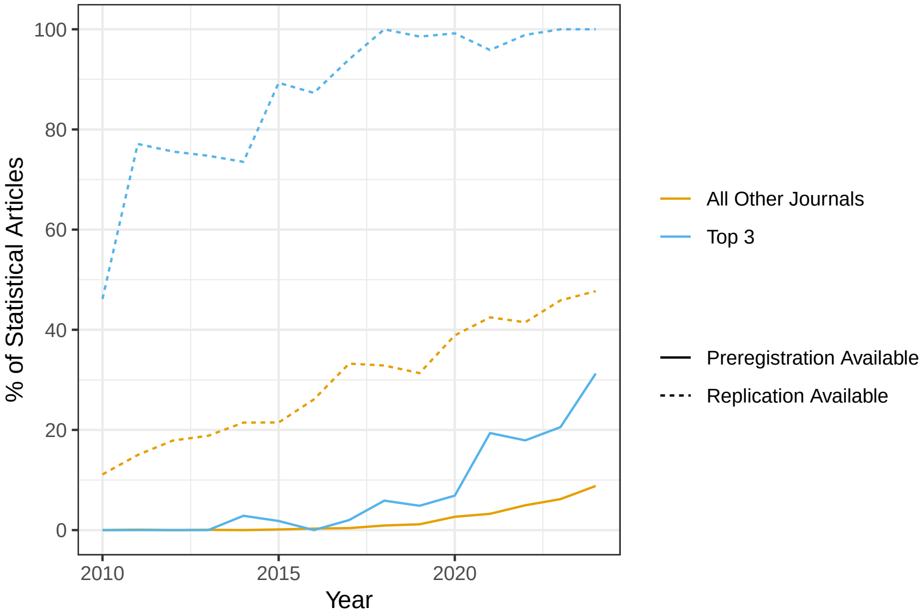

Figure 4: Share of statistical articles with available pre-registration or replication files.

One might also wonder if practices adopted at least in part to combat SoS are working. In Figure 4 we show that the share of statistical articles with available pre-registration or with replication files have seen large increases over this time period and that the increases are much more pronounced in top journals. Available replication data in top journals rose from about 50% to nearly all articles by 2018. In the collection of all articles, replication package availability rose dramatically from around 10% to 50%. At top journals nearly a third of all statistical articles are now pre-registered. In the remainder of the literature, pre-registration rose from near zero in 2015 to around 7% of articles in 2023. These are large increases so it would be reasonable to expect that SoS has gone down over time or in the places where these practices are more common.

Table 2 shows our analysis of heterogeneity, or lack thereof. The first row shows our pooled results and all subsequent rows show results in subsets of the data. Throughout we see that papers with non-null results dominate the literature. This is true in "top" journals, in articles with replication code, in articles that use modern causal inference methods, in experimental articles, across subfields, and across time.

One can find small differences across some of the categories. For example, while experiments are just as likely to have at least one non-null result in an abstract as the pooled data, they are modestly more likely to also have at least one null result. Experimental work is also almost twice as likely to have a null-only abstract as the discipline in general.

The largest difference is seen with pre-registration, where the share of articles reporting only null results

Table 2: Percentage of statistical articles published since 2010 that include null, non-null, null-only, or that reject meaningful effect sizes in their abstracts.

	Either			Only		
Group	Null	Non-Null	Reject Meaningful	Null	Non-Null	N
All journals	18	94	0.42	1.7	78	39,820
Top 3 journals	19	94	0.81	2.3	77	2,458
Preregistered	39	93	0.64	6.3	61	1,098
Replication available	19	93	0.5	2.2	76	13,212
Causal Inference	21	96	0.53	2.3	77	4,694
Experiment	28	95	0.44	3.1	70	5,227
Subfield: AP	21	94	0.74	2.8	76	6,923
Subfield: CP	18	94	0.33	1.5	78	19,936
Subfield: IR	17	94	0.33	1.5	78	4,875
Years 2010–2014	17	92	0.43	1.8	77	8,392
Years 2015–2019	17	94	0.41	1.6	78	14,660
Years 2020–2024	19	95	0.42	1.7	78	16,768

is over three times higher than in non-pre-registered articles. Pre-registered research is also 20 percentage points more likely to have at least one null. However, even in pre-registered articles, non-null results dominate.

We find a similar pattern when we disaggregate over journals. Figure 5 shows the journal-level distribution of types of claims. The left panel shows the share of articles in each journal that report at least one null result, non-null result, or claim to reject meaningful effects. The right panel shows the share of articles in each journal that *only* note either null or non-null results. Most journals report almost exclusively significant results in their abstracts.

The journal in our dataset with the largest share of articles making null only claims is the *Journal of Experimental Political Science*, where 13% of all articles are null only. *JEPS* also has the lowest share of articles making only non-null claims, at 55%. This single journal, which first published in 2014 and contributed only 234 articles to our dataset, is nevertheless responsible for 5% of the null only articles in our dataset. *JEPS* is an outlier, and all other journals show strong evidence of selection on significance. For instance, 11 journals in our dataset published at least 100 statistical articles and yet published either 1 or 0 null-only abstracts.

Taken together, Table 2 and Figure 5 point to a pattern of broad stability across political science. While certain research practices—especially pre-registration and to a lesser extent experiments—are associated with slightly better reporting profiles, these differences more-or-less amount to froth on the surface of

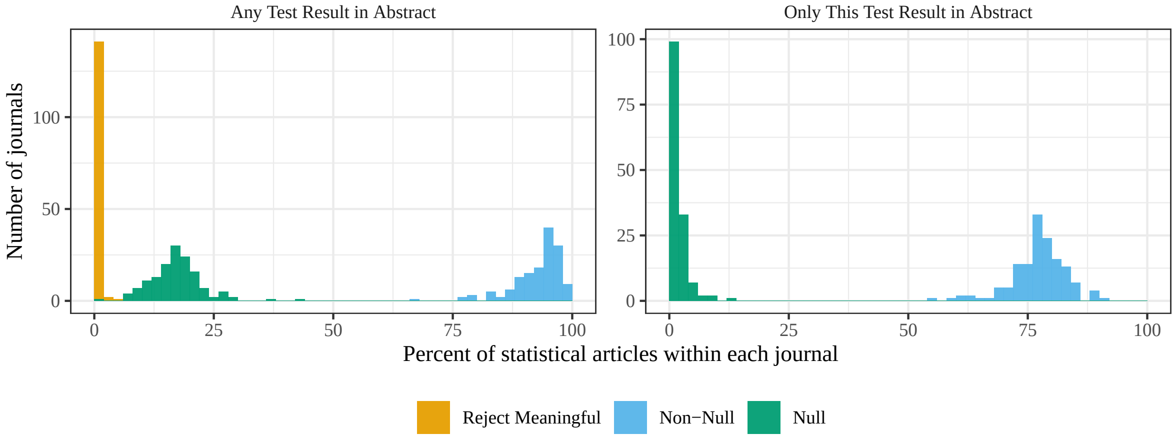

Figure 5: Journal-level null reporting

a tidal wave. Every observable feature that we can split results over produces patterns that are only explicable with very strong SoS, and differences across contextual factors are small.

Importantly, all of these results are descriptive. They allow us to identify areas of the discipline where (for whatever reason) SoS is lower, but one should not necessarily expect that if more people are compelled to use pre-registration, for example, that SoS will decrease. It is entirely possible that authors that up to now have chosen to pre-register (or publish in *JEPS*) are different in unobserved ways that also make them more likely to report null results.

7 LARGE LANGUAGE MODELS FOR META-SCIENCE: A NEW DATASET

The dataset we have created is one of the largest and most comprehensive datasets of political science articles ever created. It covers nearly all peer-reviewed political science articles published since 2010 in over 150 journals and it codes each article for 30 different methods, the presence of open science practices, the countries studied and the start and end year of any data used in each article. We expect this data to be widely useful.

To give a brief indication of the data and its uses, we describe a few fields not elsewhere discussed and show a small number of analyses. First, we record the information about the data used in each article, including start and end years and countries studied. In general, rich industrialized Western democracies are represented roughly proportional to their population relative to the United States, with some smaller data-rich countries actually over-represented relative to population (see Figure 6). For instance, the United States is studied in 5.2 times as many articles as Denmark, but has a population 56 times larger. Coverage of poorer non-Western countries is far sparser. India is covered in just 13624 articles: 3.8 times less than the United States despite having a population 4.3 times larger.

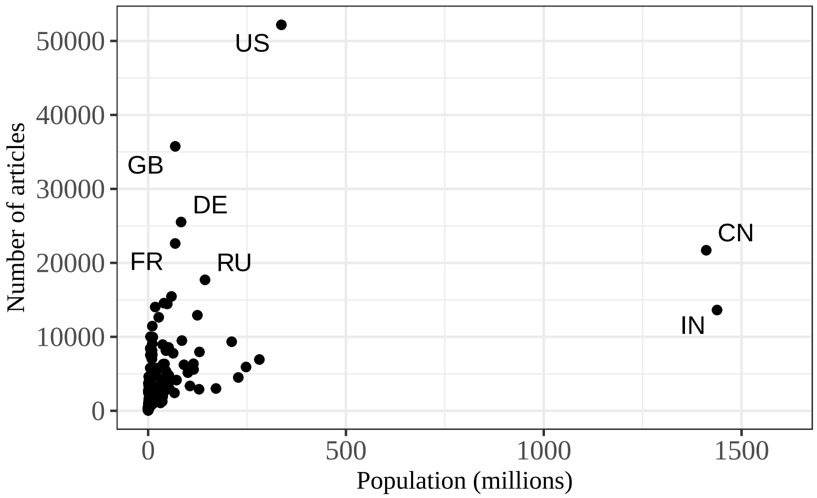

Figure 6: Relationship between a country's population and the number of times it appears in political science articles.

Second, we code information about the presence of many qualitative features like the use of historical narrative accounts, whether a paper did archival research, whether a paper took an interpretive stance, whether a paper contained an ethnography, and whether it contained a comparative case study. We code information about the presence of various observational causal inference methods (RDD, IV, synthetic control, DID) and experimental methods (audit, field, lab, survey) in papers. We also code numerous other quantitative fields like the use of regression, quantitative text analysis, and Bayesian statistics. In all we code 30 distinct fields. One can use these to track the frequency of various methods over time, an example of which is shown in Figure 7.

For the purposes of the present paper, the main result from Figure 7 is that equivalence testing is vanishingly rare, though perhaps it is growing at top journals. One might also note that in the bulk of journals, regression is now more common than historical narrative accounts and that in top journals survey experiments have rapidly grown and now also exceed historical narrative accounts in frequency.

Third, our dataset can easily enable targeted searches and follow-ups. For example, it can be used to identify everyone who wrote an article using data from only Ghana or everyone who used synthetic control methods or who published pre-registered research. This could be useful for people running a survey, for literature searches, or for various kinds of meta-science.

Fourth and finally, our dataset includes DOI identifiers for all articles. This means that with a single merge one can link the data to citation information to see how methods, topics, and open science practices relate to citation counts. It can be linked to Altmetric policy data to see which kinds of studies get referenced in policy documents. It can be linked to information about authors, open access status, or Wikipedia citations. Country-level data can be linked to country-level data on economic development, democracy, or conflict. We believe that this dataset has many potential applications.

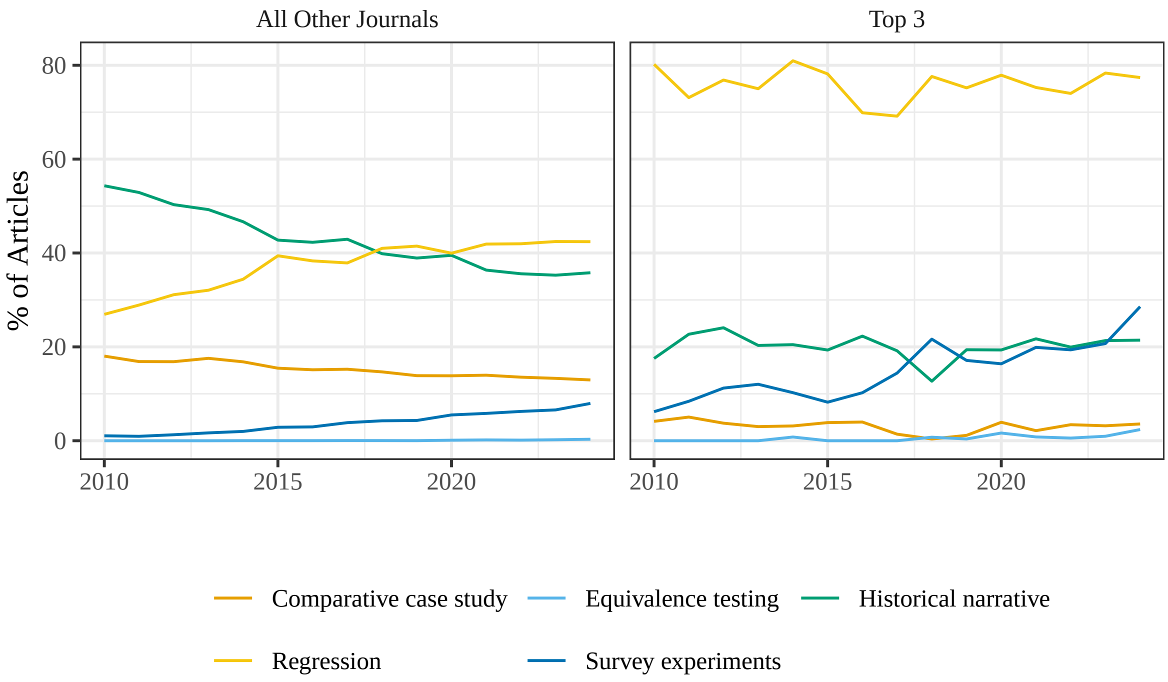

Figure 7: Frequency of various methods in journal articles.

8 CONCLUSION

Since 2010, political science has become a more statistical discipline and has warmed to replication and pre-registration. During this same time, a vanishingly small proportion of papers report only null results whereas a sizable majority report non-null results. This is hard to square with recent work showing that most papers in the discipline are greatly underpowered (Arel-Bundock et al. 2026) unless one accepts that our discipline heavily selects on statistical significance. Using a simple accounting framework, we estimate that statistically significant results are likely to be published one to two orders of magnitude more often than null results. All reasonable sets of modeling assumptions produce a striking degree of selection on significance, and this suggests that political science researchers and journal editors should redouble their efforts to reduce this bias. This effort can and should happen everywhere, as we have shown that this problem exists in all corners of the discipline.

This paper has also shown modeled LLM validation approaches that improve on current practice. We used these approaches to show how effective LLMs can be at extracting data from highly complex documents such as academic papers. With careful prompting and data preparation, current LLMs can recover ground truth data with high accuracy and consistency, often matching or exceeding highly trained graduate RAs in this use case. They do this at a fraction of the cost and time required by people. At least for this use case, the question is no longer whether it is worth sacrificing performance to code more papers at scale and lower cost. With iterative prompt development and validation, the automated approach appears increasingly competitive with even highly trained humans in this setting. We used this approach to code nearly the entire political science literature since 2010, creating a dataset that we hope will be a broadly useful public good for the discipline.

9 REFERENCES

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94. https://doi.org/10.1257/aer.20180310.

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and TD Stanley. 2026. "Quantitative Political Science Research Is Greatly Underpowered." *Journal of Politics* 88 (1): 36–46. https://doi.org/10.1086/734279.

Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. "Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale." *Proceedings of the National Academy of Sciences* 120 (41). https://doi.org/10.1073/pnas.2311627120.

Artificial Analysis. 2025. "AI Model & API Providers Analysis: Models." Artificial Analysis. 2025. https://artificialanalysis.ai/models.

Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2024. "Replication for Language Models: Problems, Principles, and Best Practice for Political Science." *Working Paper*. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf.

Berinsky, Adam J., James N. Druckman, and Teppei Yamamoto. 2021. "Publication Biases in Replication Studies." *Political Analysis* 29 (3): 370–84.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32 (4): 401–16. https://doi.org/10.1017/pan.2024.5.

Brodeur, Abel, Scott Carrell, David Figlio, and Lester Lusher. 2023. "Unpacking p-Hacking and Publication Bias." *American Economic Review* 113 (11): 2974–3002. https://doi.org/10.1257/aer.20210795.

Brodeur, Abel, Nikolai M Cook, Jonathan S Hartley, and Anthony Heyes. 2024. "Do Preregistration and Preanalysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement." *Journal of Political Economy Microeconomics* 2 (3): 527–61.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60.

Brodeur, Abel, Mathias Lemoine, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32. https://doi.org/10.1257/app.20150044.

Brodeur, Abel, David Valenta, Alexandru Marcoci, Derek Mikola, Juan P. Aparicio, Bruno Barbarioli, Rohan Alexander, et al. 2025. "Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science." I4R Discussion Paper Series, No. 195. Institute for Replication (I4R), RWI – Leibniz Institute for Economic Research.

Chambers, Christopher D, and Loukia Tzavella. 2022. "The Past, Present and Future of Registered Reports." *Nature Human Behaviour* 6 (1): 29–42.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–5.

———. 2015. "Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results." *Political Analysis* 23 (2): 306–12.

Garg, Prashant, and Thiemo Fetzer. 2025. "Causal Claims in Economics." https://doi.org/10.48550/A

RXIV.2501.06873.

Gelman, Andrew, and John Carlin. 2014. “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.” *Perspectives on Psychological Science* 9 (6): 641–51.

Gelman, Andrew, and Eric Loken. 2013. “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No Fishing Expedition or p-Hacking and the Research Hypothesis Was Posited Ahead of Time.” *Unpublished* 348 (1-17): 3.

Gelman, Andrew, and Francis Tuerlinckx. 2000. “Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures.” *Computational Statistics* 15 (3): 373–90.

Gerber, Alan, and Neil Malhotra. 2008. “Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals.” *Quarterly Journal of Political Science* 3 (3): 313–26.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. “ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks.” *Proceedings of the National Academy of Sciences* 120 (30). https://doi.org/10.1073/pnas.2305016120.

Grossman, Guy, William Dinneen, and Carolina Torreblanca. 2025. “The Evolving Landscape of Political Science: Two Decades of Scholarship in a Growing Discipline.” http://dx.doi.org/10.31219/osf.io/tmy37_v2.

Grossman, Guy, Carolina Torreblanca, William Dinneen, and Yiqing Xu. 2024. “The Credibility Revolution in Political Science.”

Hedges, Larry V. 1992. “Modeling Publication Selection Effects in Meta-Analysis.” *Statistical Science* 7 (2): 246–55.

Heyde, Leah von der, Anna-Carolina Haensch, and Alexander Wenz. 2024. “Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion.” https://doi.org/10.48550/ARXIV.2407.08563.

Huang, Fan, Haewoon Kwak, and Jisun An. 2023. “Is ChatGPT Better Than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech.” *Companion Proceedings of the ACM Web Conference 2023*, April, 294–97. https://doi.org/10.1145/3543873.3587368.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Ioannidis, John, T.D. Stanley, and Hristos Doucouliagos. 2017. “The Power of Bias in Economics Research.” *The Economic Journal* 127 (605): F236–65.

Irsova, Zuzana, Hristos Doucouliagos, Tomas Havranek, and T. D. Stanley. 2023. “Meta-Analysis of Social Science Research: A Practitioner’s Guide.” *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12595.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. “Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale.” *Research & Politics* 11 (1). https://doi.org/10.1177/20531680241231468.

Moniz, Philip, James N Druckman, and Jeremy Freese. 2025. “The File Drawer Problem in Social Science Survey Experiments.” *Proceedings of the National Academy of Sciences* 122 (12): e2426937122.

Nicholas, David, Paul Huntington, and Hamid R Jamali. 2007. “The Use, Users, and Role of Abstracts in the Digital Scholarly Environment.” *The Journal of Academic Librarianship* 33 (4): 446–53.

Ofosu, George K, and Daniel N Posner. 2023. “Pre-Analysis Plans: An Early Stocktaking.” *Perspectives* *on Politics* 21 (1): 174–90.

Palmer, Carole L, Lauren C Teffeau, and Carrie M Pirmann. 2009. "Scholarly Information Practices in the Online Environment." *Report Commissioned by OCLC Research*. https://www.oclc.org/content/dam/research/publications/library/2009/2009-02.pdf.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–41. https://doi.org/10.1037/0033-2909.86.3.638.

Scheel, Anne M, Mitchell RMJ Schijen, and Danië l Lakens. 2021. "An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports." *Advances in Methods and Practices in Psychological Science* 4 (2): 25152459211007467.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Spirling, Arthur. 2023. "Why Open-Source Generative AI Models Are an Ethical Way Forward for Science." *Nature* 616 (7957): 413–13. https://doi.org/10.1038/d41586-023-01295-4.

Sterling, Theodore D. 1959. "Publication Decisions and the Possible Effects on Inferences Drawn from Tests of Significance." *Journal of the American Statistical Association* 54 (285): 30–34. https://doi.org/10.1080/01621459.1959.10501497.

Van Zwet, Erik W, and Eric A Cator. 2021. "The Significance Filter, the Winner's Curse and the Need to Shrink." *Statistica Neerlandica* 75 (4): 437–52.

Vasishth, Shravan, and Andrew Gelman. 2017. "The Statistical Significance Filter Leads to Overoptimistic Expectations of Replicability." *Journal of Memory and Language* 95: 151–75. https://doi.org/10.1016/j.jml.2016.06.001.

Velez, Patrick, Yamil Ricardo And Liu. 2024. "Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments." *American Political Science Review*, August, 1–18. https://doi.org/10.1017/s0003055424000819.

Vivalt, Eva. 2019. "Specification Searching and Significance Inflation Across Time, Methods and Disciplines." *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816.

White, Colin, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, et al. 2025. "LiveBench: A Challenging, Contamination-Limited LLM Benchmark." In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=sKYHBTAxVa.

Yang, Eddie, Zoey Wang, Carl Zhou, and Yaosheng Xu. 2025. "Data Annotation with Large Language Models: Lessons from a Large Empirical Evaluation." Working Paper; December.

A Deviations from pre-registration

A.1 Population and sampling

Our goal in creating the population is to create a set of all original research articles published in the top political science journals from 2010 to 2024. In practice, we start with a list of DOIs from each journal and then filter out DOIs that do not correspond to original research articles. This is not always a straightforward process. Our pre-registration document noted that "Our goal is to identify the population of "original research articles," excluding published units such as book reviews, corrigenda, editorials, opinion pieces, election reports, etc." It then described a series of steps that we would take to filter out such non-original research articles from the population.

After that filtering step, we planned to randomly sample 3 articles from each journal, spread out equally over time, to create our validation sample. We expected our filtering process to have false negatives (articles that were out of scope but were marked as in scope), and so we pre-registered that "When a text is found to be out of scope by a coder (human or LLM), we exclude that text, and insert a replacement in the sample to ensure that there are 3 original research articles per journal."

A.1.1 Out of scope deviation

However, during the validation process we found that some journals published DOIs that were overwhelmingly not original research articles. We followed our pre-registered plan for 4 additional redraw rounds, each time replacing out-of-scope articles with new randomly sampled articles from the same journal in the same time period. When no in-scope article was found after these redraws, we stopped and did not reach the intended 3 articles for that journal. This occurred for 7 journals.

A.1.2 Word count deviation

Our goal was to acquire a full-text, machine readable version of the full article text (and tables) corresponding to each in-scope DOI. However, in a small number of cases it seems that the version our libraries could acquire was not the full text but just a stub. We filtered out from the population and validation sample any DOI that corresponded to an article with fewer than 2000 words. This also should remove small editorial announcements, etc., which is in line with our pre-registration.

B LLM Validation Appendix

This appendix provides additional information about our data and validation details for categorical and continuous variables, including full descriptive plots and the Bayesian performance model specification.

B.1 Journal Selection Criteria

The list of journals in our population comes from two sources. First, from Clarivate's Journal Citation Reports (downloaded in February 2025), we included journals that ranked in the top third according to either the Journal Impact Factor (JIF) or total citations, focusing on the categories of Political Science, International Relations, and Public Administration. Second, from Google Scholar Metrics (downloaded in January 2025), we took the top 20 journals in the categories of Political Science, Diplomacy and International Relations, and Public Policy and Administration. These two lists were merged and duplicates were removed.

A research assistant identified journals that are not peer-reviewed, and one of the principal investigators double-checked that coding before selecting the subset of peer-reviewed journals. Journals from two publishers were removed because of concerns raised on Beall's List and Wikipedia about academic integrity or publication practices.

B.2 Producing the Ground Truth Data

To test human and LLM performance, we created a validation sample by drawing one article per journal per 5-year period from 2010–2024. Each paper in the validation sample is coded by two humans and three LLMs: GPT-5, GPT-5 mini, and Kimi K2. Testing prior to validation suggested that GPT-5 was strong on this task, and so any disagreement between the humans and GPT-5 was then reconciled by a professor in a separate interface. If both humans and GPT-5 agreed on a field then that value is treated as correct. These correct answers were then used to score the human coders and each of the three LLMs.

The LLM coding was done via API requests to OpenAI using the Langchain framework with structured generation to ensure a specific output format. Our main LLM for coding the validation sample was GPT-5 2025-08-07. We also tested the performance of GPT-5 mini 2025-08-07, a substantially cheaper model than GPT-5, and Kimi K2 Thinking, which was plausibly the best open weights model at the time of testing.

For each paper in the validation sample, each coder first identified whether the paper is within scope for our project (whether it was substantive research). This is necessary because we were unable to perfectly filter book reviews or editorial letters, for example, from the population of articles. When a text is found to be out of scope by a coder (human or LLM), one of the principal investigators reviewed that decision and excluded the text if appropriate. Following our pre-registration plan, we then inserted a replacement article drawn at random from the same journal-time period. Articles from the validation sample were presented to human coders in random order, until two human coders completed the coding task for each article.

If there are discrepancies between any of the human coders or any human and GPT-5 then the results were reconciled by one of the principal investigators of this project (with further team discussion on any hard cases). Reconciliation was done blindly, without knowing if an LLM or human made a certain choice. The end result of this process is a ground truth dataset that is formed from a mix of complete agreement across coders and/or reconciliation by a principal investigator. This has the advantage of using PI time efficiently (only examining the hard cases) and giving focused attention to the places where there is disagreement among coders. The main downside to this approach is that it will create incorrect gold standard data in the case where all coders agree on the wrong answer.

The final validation dataset includes 701 articles coded by research assistants and 644 coded by the LLMs. The reason why there are more articles coded by research assistants is that we were unable to sign contracts with all publishers to obtain XML files and sometimes HTML collection was quite challenging. This was not an issue for the humans as they coded articles based on the PDF versions of the articles from the publishers' websites.

To compare the performance of LLMs and human coders we consider four metrics. Accuracy measures the overall proportion of correctly classified cases out of all cases. Sensitivity (or recall) focuses on the share of true positives correctly identified among all actual positives. Specificity captures the proportion of true negatives correctly identified among all actual negatives. Precision evaluates how many of the predicted positives are in fact correct.

We have three contrasts that can be examined for any paper-level variables. First, and most critically, we evaluate how the LLM performs at recovering the ground truth data. Second, we see how the humans perform at this same task. Third, we can compare the performance of the LLM at recovering ground truth against that of the humans. We designed our validation setup to ensure sufficient statistical power to draw robust conclusions about the first of these three questions.

B.3 Unmodeled Coder Performance

Figure 8 shows the performance metrics for our three LLMs. In all four plots, dots further to the right indicate better performance and each row is a feature of an article that was coded. Kimi K2 consistently performs worse than the Open AI models, while GPT 5 mini seems to do about as well as GPT 5.

Figure 9 compares the performance of GPT-5 mini (orange dots) to the human coder average (blue triangles) across these four metrics. The LLM performs well overall, with most of their low scores being on rare (noisy) fields. GPT 5 mini usually beats the humans on raw accuracy, in some cases by a significant margin. It also does quite well on sensitivity, or correctly coding that a feature exists. This is especially important in our context, as a central part of the task is reading a complex document and identifying when a rare feature exists.

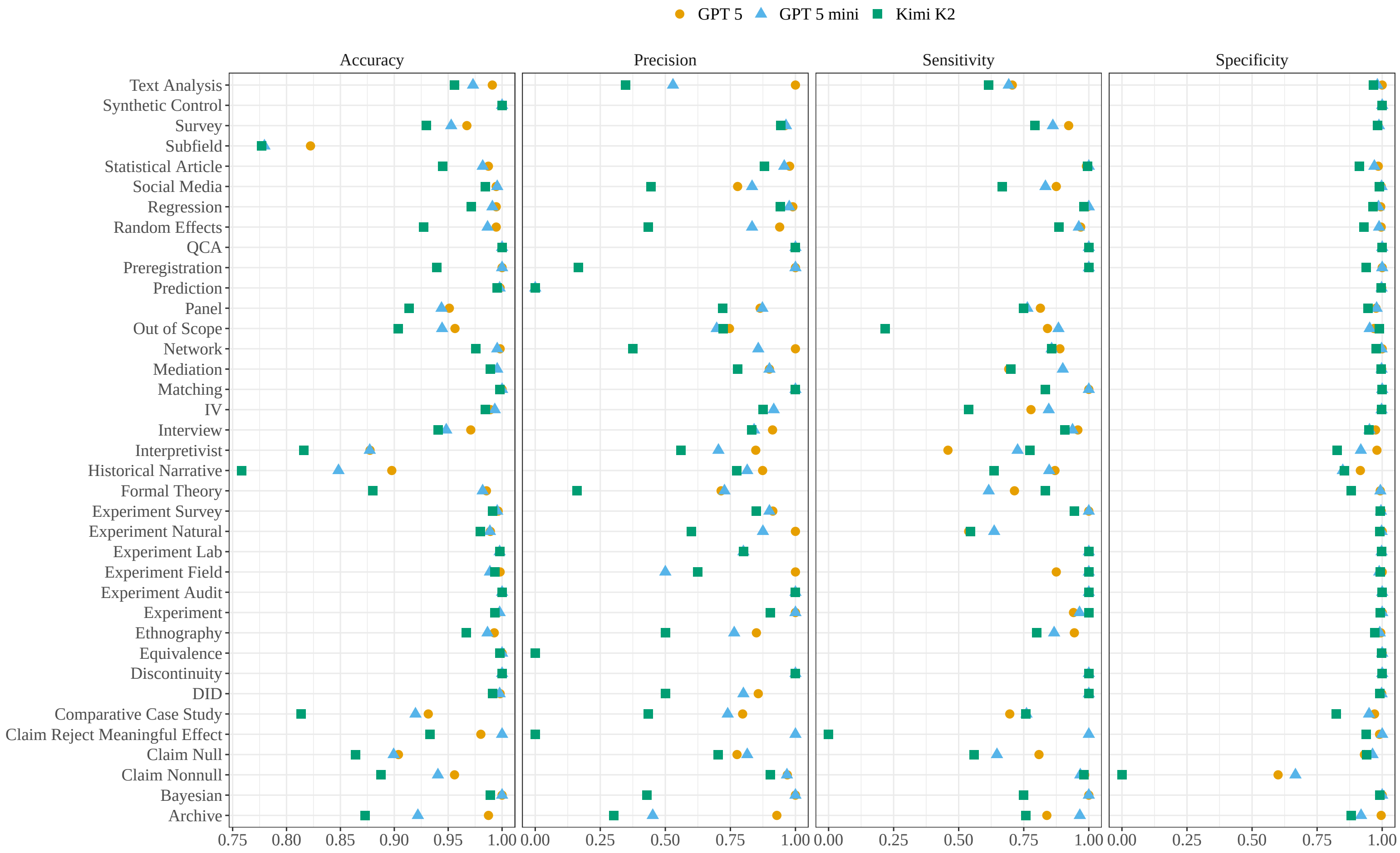

Figure 8: LLM vs. LLM (unmodeled)

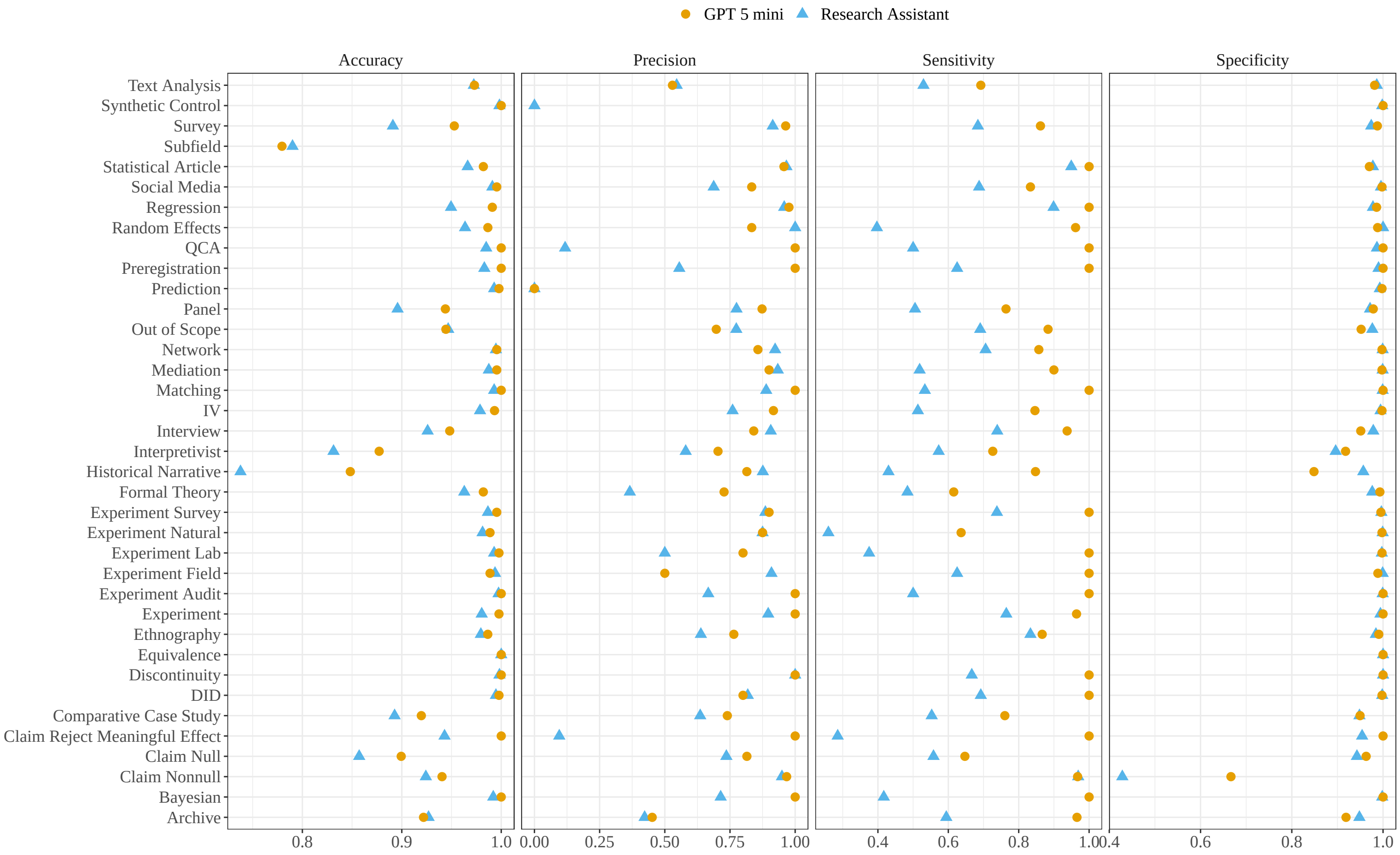

Figure 9: LLM vs. Human (unmodeled)

Table 3: Accuracy for years

	Mean absolute error		% with errors $>$ 4 years	
	Start year	End year	Start year	End year
GPT 5 2025-08-07	2.5	0.4	6.6	2.0
GPT 5 mini 2025-08-07	3.3	0.5	9.4	2.0
Kimi K2 thinking	8.9	8.2	20.0	6.3
Research Assistants	14.8	22.5	10.6	3.6
	% Wrongly skipped		% Wrongly added	
	Start year	End year	Start year	End year
GPT 5 2025-08-07	2.6	2.6	0.5	0.5
GPT 5 mini 2025-08-07	3.1	3.1	0.4	0.4
Kimi K2 thinking	6.4	5.8	1.3	1.3
Research Assistants	6.8	7.8	1.3	1.2

B.3.1 Continuous Feature Performance

So far, we have only evaluated performance when coding categorical variables, but we also asked coders to extract the start and end years for any data used in any article. For these continuous features we measure accuracy using mean absolute error (MAE), the fraction of cases that missed by over 5 years, the fraction of cases where the ground truth had a numeric answer but the coder did not provide one, and the fraction of cases where the coder provided a numeric answer but the ground truth did not.

Similarly to the factor-level features, we find that LLM performance is strong. GPT 5 fairly consistently scores best, with GPT 5 mini a close second. Kimi K2 and the human average have both much higher rates of errors and more severe errors. They are also more likely to wrongly skip and wrongly add year information.

B.4 Modeled Coder Performance

For our main validation procedure we estimate a model where the coder answer (y_i) is modeled as the outcome of a Bernoulli distribution with probability p_i where i is a measure, j is a paper, and k is a coder.

$$y_{ijk} \sim \text{Bernoulli}(p_{ijk}),$$

This probability is then a function of the predicted sensitivity (η_{ijk}^{sens}) and specificity (η_{ijk}^{spec}). The relevant prediction is determined by whether the true value for this variable (t_{ijk}) is true or false. Essentially, the interaction terms mean that the sensitivity and specificity are from two separate equations. However, as we will show next, we allow some covariance in random effects across these questions.

$$p_{ijk} = t_{ij}\,\sigma(\eta^{\text{sens}}_{ijk}) + (1 - t_{ij})\,[1 - \sigma(\eta^{\text{spec}}_{ijk})].$$

We also include random slopes for measures (α_i^{sens}, α_i^{spec}), and coders ((α_k^{sens}, α_k^{spec})), which are further nested within whether the coder is human ($\alpha_{H_k}^{sens}$, $\alpha_{H_k}^{spec}$).

$$\eta_{ijk}^{sens} = \alpha^{sens} + \beta_{prev}^{sens}\,prevalence_i + \beta_{human}^{sens}\,H_k + \alpha_i^{sens} + \alpha_k^{sens} + \alpha_{H_k}^{sens}$$

$$\eta_{ijk}^{spec} = \alpha^{spec} + \beta_{prev}^{spec}\,prevalence_i + \beta_{human}^{spec}\,H_k + \alpha_i^{spec} + \alpha_k^{spec} + \alpha_{H_k}^{spec}$$

All the random intercepts are allowed to covary according to a correlation matrix with a $LKJ(1)$ prior.

B.4.1 Modeled Results

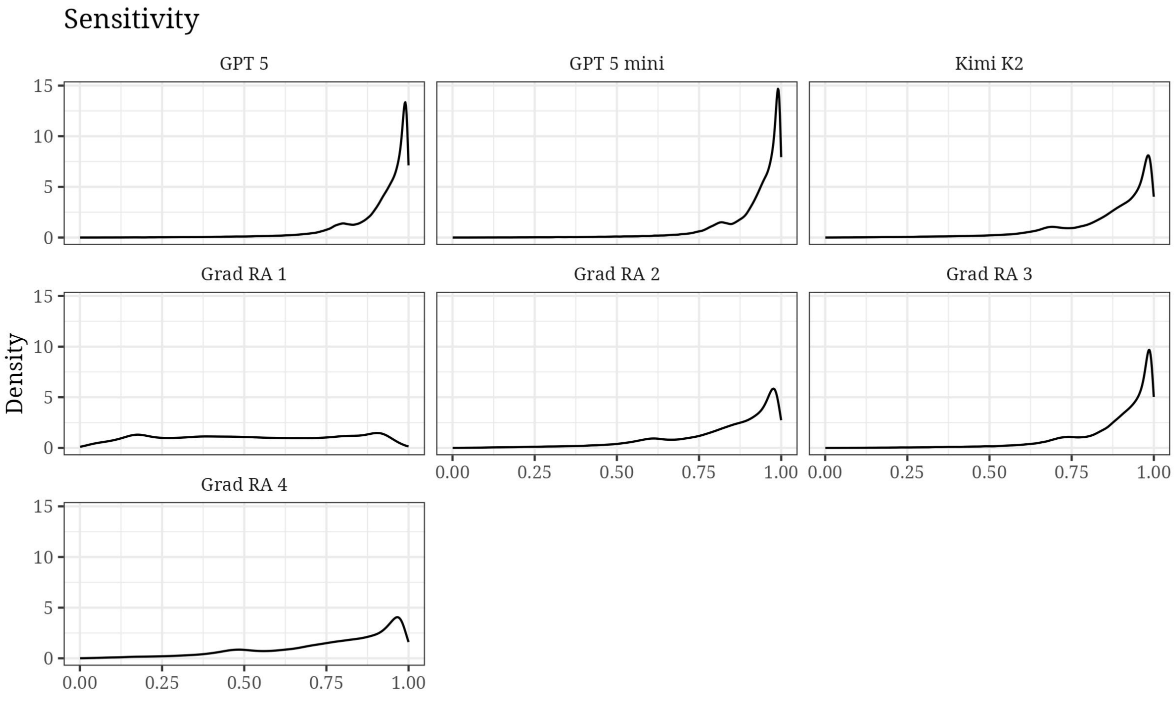

Figure 10: Modeled individual coder sensitivity across all features

Figure 10 unpacks aggregated differences in sensitivity by individual coder. LLM coders exhibit little variation in sensitivity and specificity across variables, though Kimi K2 does worse than the other LLMs. The human coders show greater heterogeneity. While no human does better than the Open AI models, one comes close to Kimi K2 and a single low-performing coder does much worse than the rest, reducing aggregated human accuracy. In sum, the best human coders perform slightly worse than the worst LLM on sensitivity, though human performance on specificity is more comparable.

C Paper-Level Coding Guidance for RAs

This document describes how to turn articles into structured data using our web interfaces. This is the second version of the document. The human and LLM codebooks are largely identical, with only minor format and feature differences.[14]

- The paper coding interface is here: **URL**

Access the interface using Google Chrome. When you open the interface, you will sign in with your email address. Always use your university email address. When the interface loads you will see the title of an academic article and a link to it. Load the article (you may need to access it through your library if you are not on campus internet or on a VPN) and then fill out the fields. Add each table using the "Add New Table" button. When everything looks right, click "Submit" to get a new article.

Why are we doing this?

Our project is exploring using large language models (LLMs) to extract information from academic articles at scale. If we can do this, then we will be able to very quickly and cheaply turn an academic article into machine readable data. This is very useful for a range of meta-science topics.

From our pilot testing, it seems like we can get LLMs to do this reasonably well. However, we need to more carefully estimate LLMs performance relative to humans or relative to gold-standard data. You are helping us create both the human-coded data and the gold-standard data against which we will test model performance. You are not generating data that will be used to train models, as we are running inference on the LLMs but not training them.

Each article will be coded by at least two people and by the LLM. It is important that you code each article independently (do not ask each other for help), as we are using your answers to understand how often human coders agree on the fields. We expect some level of disagreement, both because there will be some ambiguity in the coding rules and due to normal human error. After two people have coded an article, a professor involved in the project will review any disagreements between coders while being blinded to who gave what answers. We will resolve all disagreements and this will produce our gold standard dataset. We are aiming to do this for about a few hundred papers. We will start with article and table-level information and if this goes well we will later circle back to each paper and extract information about models and coefficients.

[14] The main difference between the human and LLM codebooks is that the human codebook sometimes has more involved formatting like bullets for listing items, the LLM codebook sometimes has words in ALL CAPS, and the LLM codebook has a small number of additional features that we did not ask the humans to code because of resource constraints.

Descriptions of study-level fields

The following fields are article-level, so there will be one answer per article.

For all fields, answer "uncertain" if any field cannot be determined confidently from the article. In general, we would like you to take your time and think carefully about a choice rather than moving quickly to selecting "uncertain."

Out of scope

We are only interested in articles that present original research, so an article is out of scope if it does not present original research. For example, book reviews, comments from editors, or tables of contents may receive a DOI and so might appear in our dataset but we have no interest in them. They should be marked as "Yes." Review articles—which mainly summarize a body of literature—are out of scope. Most articles will be in scope. An article can present original research even if it is not empirical. For example, political theory articles that develop original arguments are in scope.

Statistical paper

Indicate whether the study uses statistical methods to derive any of the results in the main article (i.e. ignore appendices). Statistical methods are methods that quantify uncertainty (standard errors, p-values, confidence intervals, significance indicators, t-statistics, etc.) in empirical estimates (i.e. exclude articles that only report point estimates without any indication of uncertainty). Do not include studies that use only simulation or formal modeling.

Pre-registration

Was the study pre-registered? If there is no indication that the study was pre-registered, select "no". Use the response "uncertain" only if there is genuine uncertainty (i.e. mixed signals) about whether the study was pre-registered. If the study was pre-registered then give the URL of the pre-registration document for this study in the text box.

Claims non-null findings in abstract

A null result is one where the authors characterize a finding as not supporting the existence of a phenomenon in terms of failing to find evidence for a phenomenon. A non-null result is one where authors characterize a finding as supporting the existence of a phenomenon by rejecting the null hypothesis.

Indicate "yes" if the abstract mentions non-null results (supporting the existence of a phenomenon), "no" if no non-null results are mentioned, or "uncertain" if it is unclear.

Claims null findings in abstract

A null result is one where the authors characterize a finding as not supporting the existence of a phenomenon in terms of failing to find evidence for a phenomenon. A non-null result is one where authors characterize a finding as supporting the existence of a phenomenon by rejecting the null hypothesis.

Indicate "yes" if the abstract mentions null results (failing to find evidence for a phenomenon), "no" if no null results are mentioned, or "uncertain" if it is unclear.

Rejects meaningful effects in abstract

Mark "yes" if the abstract states that the study's results rule out effects that are meaningful based on theory or practical considerations. This includes studies claiming a "precise zero effect." Mark "no" if the study does not make such a claim. Mark "uncertain" if it's unclear whether the study rules out meaningful effects.

Replication available

Please answer "available" if the code, data, or replication package is stated as being publicly available in the text of the paper. Answer "on request" if the authors state that the code, data, or replication package is available upon request or similar. Answer "none" if there is no indication of availability. Answer "uncertain" if you are not sure. Please paste the url to the code/data in the text box if it is available.

Methods

Please answer a series of "yes", "no", or "uncertain" questions about the methods used in the article. An article should be coded as "Yes" if it is unambiguous that it uses the technique. Choose "Yes" if the article clearly describes the method in a way that makes clear that it is one of the methods used in the article. Also answer "Yes" if the model formula implies that a given method is used or if the name of the method is explicitly used.

- **Formal theory**, game theory, or other mathematical theoretical methods. Statistical methods and simulation-based methods do not count here.
- **Archives** is research that uses primary sources such as historical documents. It is typically conducted in archives or libraries. Only code this as yes if the author or research assistants extracted primary source information from an archive while doing their research.
- The **comparative case study** approach aims at deducing cause and effect via comparisons across cases. Cases are typically referred to as such and should have clear temporal and spatial bounds. Examples include: Mill's methods, small-N comparative case studies, most-similar or most-different cases. This does not include large-N statistical analyses.

- **Historical narrative** are papers that present a qualitative historical narrative account as part of its contribution or results. Historical in this context includes recent history. Code this as no if such a narrative is present but merely used as background or given as the motivation for another type of analysis. One common form of qualitative historical narrative is the case study.
- **Interpretivist** approaches prioritize understanding meaning, context, and subjective experiences rather than seeking generalizable laws. Indicators include: explicit discussion of meaning-making, social construction, reflexivity, subjectivity, or verstehen; reliance on hermeneutics, phenomenology, the primacy of intersubjective and lived experiences, or discourse analysis in an interpretivist framework; rejection of objectivity or positivist assumptions; or arguments emphasizing context-dependent knowledge. Only say yes if the paper frames some of its methodology within an interpretivist epistemology.
- **Ethnography** is a method involving prolonged and immersive engagement with a group of people. It typically includes field observations, participant observation, and extended interaction within a community. Select "yes" if the paper shows the researcher's sustained, on-site engagement or participant observation. Interviews may be present, but on their own they do not constitute ethnography.
- **Interviews** are verbal data collection methods directly eliciting responses from participants in a structured, semi-structured, or unstructured format (including focus groups). Interviews can be conducted in person or remotely. Select "yes" if the authors conducted interviews or focus groups. Do not include surveys or purely secondary data analysis.
- **Qualitative Comparative Analysis (QCA)** - Does the paper do qualitative comparative analysis (QCA) or otherwise combine boolean logic and set theory with a small to medium number of cases in order to answer research questions?
- **Experiment** means that the researcher controlled or worked with someone who controlled how the treatment was applied randomly. Do not include natural experiments. Randomized controlled trials (RCTs) are a type of experiment.
- A **field experiment** is a randomized experiment conducted in the field, outside the lab, in the real world.
- An **audit experiment** is a type of field experiment where experimenters contact a person or institution to request something (e.g. information, a job, a meeting, or even just a reply). The attributes of the request (e.g. what is being requested) or the requestor (e.g. their partisanship, ethnicity or gender) are varied with the goal of understanding how the people who are contacted respond to different requests and requestors.
- A **lab experiment** is a randomized experiment conducted in a laboratory setting.
- A **survey experiment** is a randomized experiment conducted in a survey.
- **Survey** research. Survey questions can be asked in person, by phone, by mail, or online. Any analysis of data from a survey counts.
- **Social media data** - Does this paper analyze data directly sourced from a social media service? Examples include data from: Twitter or X, Facebook, Reddit, Instagram, Snapchat, LinkedIn, YouTube, TikTok. Wikipedia and Google trends data does not count. Do not say yes if the paper merely has social media usage as a variable in an analysis (e.g. from a survey) but the data do not come from a social media service.
- **Panel data** marks any use of panel data, which is data with repeated observations of the same unit

(person, village, company, state, dyad) over time (day, week, year). Using a single wave of data from a panel dataset does not constitute a use of panel data (e.g. wave 2 of the British Election Study Internet Panel would not count but waves 3 and 5 of the British Election Study Internet Panel would count because it involves analysis of the same units over time).

- A **natural experiment** means that the research design exploits variation in some random or quasi-random event that occurred in nature or under the control of some entity other than the researcher. Do not include controlled experiments.
- **Network analysis** studies the relationships between entities. This will often involve terminology like nodes, edges, centrality, ties, social network analysis, ERGM, TERGM, etc.
- **Text analysis** or content analysis involves analyzing text data to identify patterns, relationships, and trends. It can focus on the frequencies of certain words, topics, or other patterns in the text. Only include quantitative text analysis.
- Does the paper estimate **random effects**? Indicators of random effects include: random intercepts, random slopes, multilevel models, hierarchical models, bayesian hierarchical model, bayesian multilevel model, or mixed effects. Only say yes if the paper estimates random effects.
- **Bayesian statistics** includes Bayesian statistical models and concepts like "markov chain monte carlo" or "posterior distribution". Only code as "yes" if Bayesian concepts are used in a statistical context. This does not include other uses of Bayesian ideas in game theory or other non-statistical contexts. The use of the Bayesian information criterion (BIC) to test model fit does not mean that an article used Bayesian statistics.
- **Difference-in-differences** Do the authors explicitly describe at least one of their analyses as a "difference in differences" (or DiD)? Triple differences also count as "yes".
- **Regression Discontinuity Design (RDD)** is a causal identification approach that relies on the assumption that all potentially relevant variables aside from the outcome and treatment are continuous at the point where a discontinuity occurs. It will usually be referred to as 'regression discontinuity" in the text.
- **Instrumental variables** are a causal identification approach that relies on an exclusion restriction. It will usually be referred to as "instrumental variables" in the text and will usually be estimated with two-stage least squares.
- **Synthetic control method** is a causal identification approach that uses panel data to build a counterfactual (or synthetic) unit via a weighted average of other units. The synthetic unit is weighted to closely approximate the unit that is eventually treated, and the difference in the trajectory of the synthetic control unit vs the real treated unit is the causal effect. If it is used one will usually find the words "synthetic control" in the text.
- **Matching** and balancing methods, such as propensity score matching, coarsened exact matching, Mahalanobis distance matching, inverse probability weighting, structural marginal models, propensity score, or entropy balancing. Do not count as "yes" methods that only use survey weights.
- **Prediction** is a kind of analysis where the goal is to minimize the prediction error, or to forecast an outcome in the future. Predictions will always be made for data points that are out of sample, like election returns in the future. Examples include multilevel regression with post-stratification (MRP), and election forecasting.
- **Regression** modelling, such as linear models, ordinary least squares, generalized linear models,

categorical outcome models, etc.

- **Equivalence tests** are formal statistical tests used to determine whether the effect of interest is small enough to be considered theoretically or practically insignificant. There are a small number of common methods to do this. Code this as "yes" only if one or more of the following methods are explicitly used: TOST (Two One-Sided Tests), Intersection-Union Test (IUT), non-superiority test, non-inferiority test, or if a paper uses Bayesian statistics then the Region of Practical Equivalence (ROPE). No other methods should be included.
- **Mediation analysis** examines whether the effect of an independent variable on a dependent variable operates through one or more mediators. The paper will explicitly use the word "mediation" to describe the analysis.

Country universe

For this field please give a short textual description of the countries covered. e.g. "India", "OECD countries", "subsaharan Africa", "former French colonies" etc.

Start year and end year

Indicate the earliest and latest year covered by the data used in empirical analysis in this article. If the article only gives a qualitative hint at the date, then use your best judgment to enter a date that you think is as close to the correct date as possible. For example, if an article on African politics says their dataset "starts at independence" then it would be reasonable to enter 1957 if you knew Ghana's independence date or 1960 if you were more crudely guessing. An article might rarely give no or next-to-no indication of the start or end years, and in that case leave these fields blank. Do not enter a zero or some other number in order to convey that you are uncertain or that the article lacks year information or data that spans time. In these situations just leave it blank.

C.1 Subfield

The subfield of political science to which an article belongs. The options for subfield are:

- American politics - questions of domestic institutions or political behaviour in the United States. For a paper to be American politics it must draw only on American evidence
- Comparative politics - questions of domestic institutions or political behaviour outside of the United States. An article that would be coded as comparative politics but only uses evidence from the USA should be coded as American Politics
- International Relations - questions related to international politics. This includes security studies and studies of civil war. Studies that link international causes to domestic outcomes, or domestic causes to international outcomes, should be counted as international relations.
- Political methodology - research primarily about methodology

- Political theory - The field of political theory has two parts: normative and positive political theory. Normative political theory is related to the field of political philosophy and always belongs in this category. Positive political theory includes formal or mathematical theory and theory expressed only in words. Positive political theory should only be coded as political theory if it does not fit well in comparative politics, American politics, international relations, or political methodology.
- Other - anything else

C.2 Comments

If you are uncertain about any of the paper-level fields, write here a brief description of your reasoning and what the other right answer might be.

D Checks of Face Validity

This appendix shows a series of non-pre-registered analyses of our full dataset. These were created to probe the face validity of our full data.

Figure 11 shows how our subfield codings vary over time within various journals. These journals were selected because we either have strong priors about which subfields should be dominant, or because we believe that they should be mixed. The results accord with our expectations.

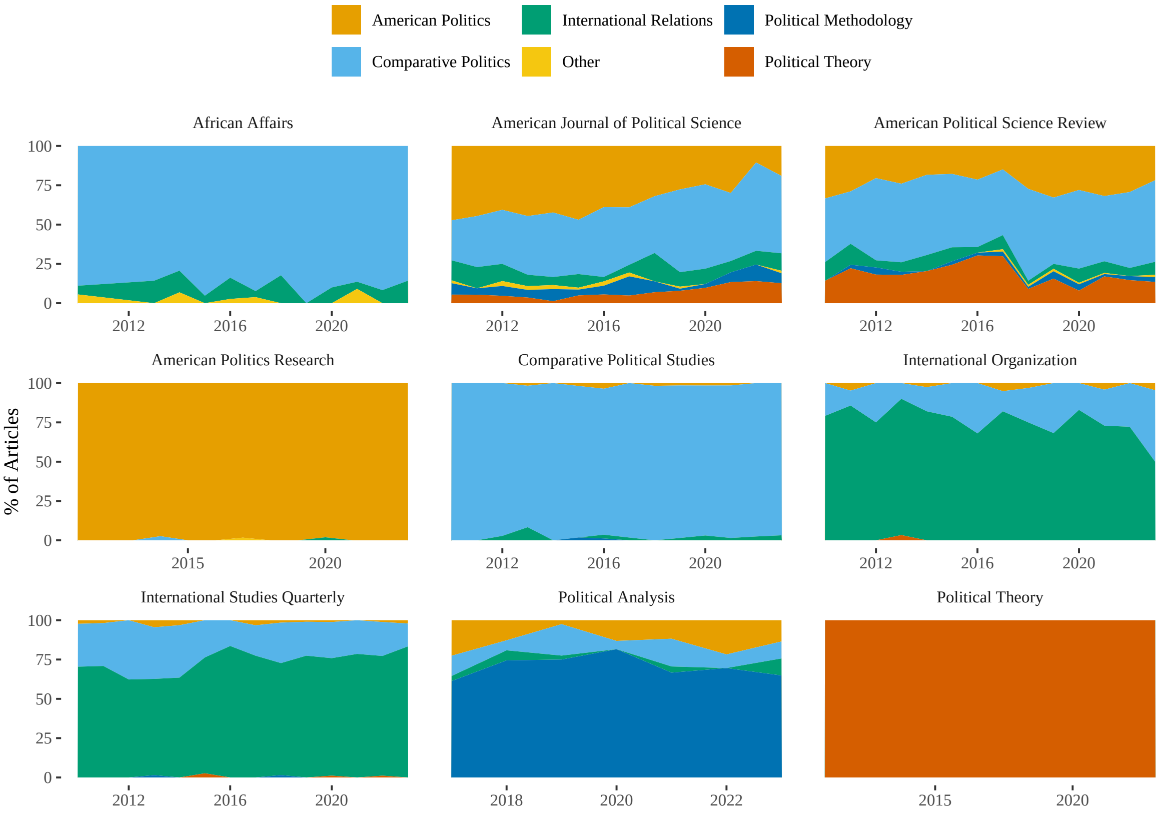

Figure 11: Subfield distributions over journals.

Figure 12 shows the rates of co-occurrence between different methods used in the papers we coded. One can see that our subtypes of experimental methods (but not natural experiments) always co-occur with the more general experimental feature. Regression is used when many of our quantitative methods are used. When someone does an ethnography they typically also do interviews. When someone does a survey experiment, they also do a survey.

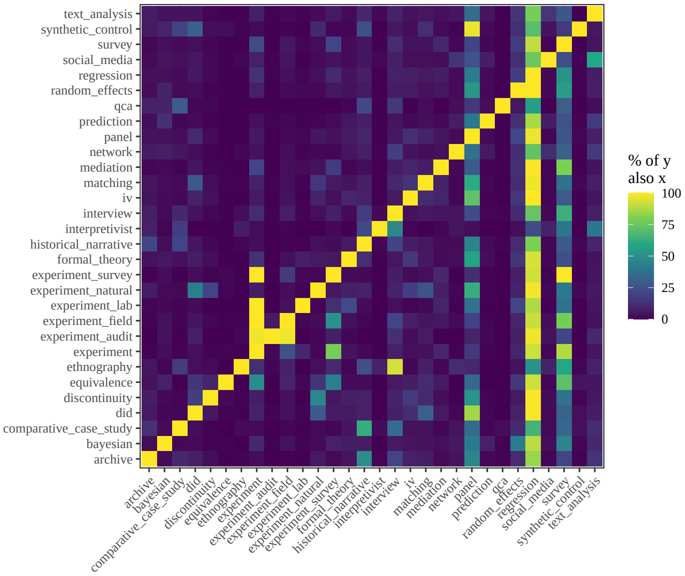

Figure 12: Co-occurrence of methods.

E Journals in our population

Table 4: Population of journals.

Journal	Years covered	N papers	N in scope
Administration and Society	2011-2024	745	666
African Affairs	2010-2024	355	303
American Journal of International Law	2017-2024	413	125
American Journal of Political Science	2010-2024	964	964
American Political Science Review	2010-2024	1117	1102
American Politics Research	2011-2024	642	642
American Review of Public Administration	2010-2024	615	563
Armed Forces and Society	2010-2024	606	532
Asia Pacific Journal of Public Administration	2014-2024	240	196
Australian Journal of International Affairs	2010-2024	645	479
Behavioural Public Policy	2017-2024	267	157
British Journal of Political Science	2010-2024	901	877

Continued on next page

Table 4: Population of journals. (Continued)

British Journal of Politics and International Relations	2010-2024	718	658
Cambridge Review of International Affairs	2010-2024	603	505
Canadian Journal of Political Science	2010-2024	630	575
Chinese Journal of International Politics	2010-2024	264	240
Chinese Political Science Review	2016-2023	181	176
Citizenship Studies	2010-2024	890	815
Climate Policy	2011-2024	1131	935
Comparative European Politics	2010-2024	433	383
Comparative Political Studies	2010-2024	967	942
Conflict Management and Peace Science	2011-2024	431	411
Contemporary Italian Politics	2013-2024	295	263
Contemporary Political Theory	2010-2024	441	322
Contemporary Politics	2010-2024	476	440
Contemporary Security Policy	2010-2024	394	354
Cooperation and Conflict	2011-2024	417	381
Critical Policy Studies	2010-2024	448	367
Critical Studies on Security	2013-2024	248	197
Democratization	2010-2024	1032	978
Earth System Governance	2019-2024	151	115
East European Politics	2012-2024	399	371
Electoral Studies	2010-2024	1592	1519
Emerging Markets Finance and Trade	2015-2024	2180	2162
Environmental Politics	2010-2024	905	800
EuropeAsia Studies	2010-2024	1119	1029
European Journal of International Law	2010-2024	943	573
European Journal of International Relations	2010-2024	582	574
European Journal of International Security	2016-2024	209	202
European Journal of Political Economy	2010-2024	1145	1128
European Journal of Political Research	2010-2024	778	722
European Political Science Review	2010-2024	431	426

Continued on next page

Table 4: Population of journals. (Continued)

European Security	2010-2024	477	452
European Union Politics	2011-2024	439	422
Foreign Policy Analysis	2010-2024	415	408
Geopolitics	2010-2024	869	737
German Politics	2010-2024	484	447
Global Public Policy and Governance	2021-2023	69	62
Globalizations	2010-2024	1145	961
Governance	2010-2024	592	559
Government Information Quarterly	2010-2024	997	928
Government and Opposition	2010-2024	510	447
Human Service Organizations Management	2014-2024	382	305
Intelligence and National Security	2010-2024	871	669
International Affairs	2013-2024	912	790
International Feminist Journal of Politics	2010-2024	516	426
International Interactions	2010-2024	566	528
International Journal	2013-2024	414	290
International Journal of Public Administration	2010-2024	1488	1391
International Journal of Public Opinion Research	2010-2022	429	410
International Journal of Transitional Justice	2010-2024	458	364
International Organization	2010-2024	461	440
International Political Science Review	2011-2024	572	526
International Political Sociology	2010-2024	390	369
International Public Management Journal	2010-2024	485	462
International Relations	2011-2024	453	394
International Review of Administrative Sciences	2011-2024	648	600
International Studies Perspectives	2010-2024	369	314
International Studies Quarterly	2010-2024	1141	1110
International Studies Review	2010-2024	625	390

Continued on next page

Table 4: Population of journals. (Continued)

International Theory	2010-2024	313	234
Journal of Accounting and Public Policy	2010-2024	576	566
Journal of Chinese Governance	2016-2024	239	223
Journal of Chinese Political Science	2010-2024	371	327
Journal of Common Market Studies	2010-2024	1281	1188
Journal of Comparative Policy Analysis Research and Practice	2010-2024	479	403
Journal of Conflict Resolution	2011-2024	923	895
Journal of Current Southeast Asian Affairs	2010-2024	294	268
Journal of Elections Public Opinion and Parties	2010-2024	480	471
Journal of European Integration	2010-2024	867	758
Journal of European Public Policy	2010-2024	1515	1393
Journal of European Social Policy	2012-2024	469	436
Journal of Experimental Political Science	2014-2024	242	237
Journal of Genocide Research	2010-2024	492	314
Journal of Information Technology Politics	2010-2024	400	381
Journal of International Relations and Development	2010-2024	461	424
Journal of Intervention and Statebuilding	2010-2024	492	383
Journal of Peace Research	2010-2024	927	886
Journal of Political Marketing	2010-2024	268	251
Journal of Political Power	2011-2024	402	269
Journal of Politics	2015-2023	1084	1074
Journal of Public Administration Research and Theory	2010-2024	604	550
Journal of Public Policy	2010-2024	352	345
Journal of Social Policy	2010-2024	633	601
Journal of Strategic Studies	2010-2024	622	526
Journal of the Japanese and International Economies	2010-2024	430	429
Latin American Politics and Society	2018-2024	215	182

Continued on next page

Table 4: Population of journals. (Continued)

Local Government Studies	2010-2024	735	688
Marine Policy	2010-2024	4829	4307
Mediterranean Politics	2010-2024	513	384
New Political Economy	2010-2024	740	690
Ocean Development International Law	2010-2024	301	286
Parliamentary Affairs	2010-2024	696	613
Party Politics	2010-2024	1025	1002
Perspectives on Politics	2010-2024	925	514
Perspectives on Public Management and Governance	2017-2024	153	103
Policy Design and Practice	2018-2024	200	167
Policy Sciences	2010-2024	373	339
Policy Studies	2010-2024	629	574
Policy and Society	2010-2024	430	377
Political Analysis	2017-2024	289	288
Political Behavior	2010-2024	664	659
Political Communication	2010-2024	514	458
Political Geography	2010-2024	1533	1184
Political Research Quarterly	2010-2024	1075	1063
Political Science Quarterly	2013-2024	289	202
Political Science Research and Methods	2013-2024	595	590
Political Studies	2010-2024	974	960
Political Theory	2011-2024	469	377
Politics	2010-2024	502	457
Politics Groups and Identities	2013-2024	658	575
Politics Religion Ideology	2011-2024	359	307
Politics and Gender	2010-2024	550	455
Politics and Religion	2010-2024	432	417
Politics and Society	2011-2024	302	267
PostSoviet Affairs	2013-2024	318	297
Public Choice	2010-2023	1052	946

Continued on next page

Table 4: Population of journals. (Continued)

Public Management Review	2010-2024	1265	1170
Public Money Management	2010-2024	1011	829
Public Opinion Quarterly	2010-2024	634	595
Public Performance Management Review	2014-2024	477	461
Public Personnel Management	2013-2024	298	285
Public Policy and Administration	2011-2024	317	283
Publius the Journal of Federalism	2010-2024	414	365
Research and Politics	2014-2024	507	503
Review of African Political Economy	2010-2024	632	405
Review of International Organizations	2010-2023	310	292
Review of International Political Economy	2010-2024	823	751
Review of International Studies	2010-2024	820	764
Review of Public Personnel Administration	2011-2024	379	359
Review of World Economics	2010-2023	385	381
Science and Public Policy	2012-2014	167	143
Security Dialogue	2011-2024	476	444
Security Studies	2010-2024	425	353
Social Movement Studies	2010-2024	652	590
South European Society and Politics	2010-2024	392	368
Space Policy	2010-2024	532	355
Studies in Comparative International Development	2010-2023	303	293
Studies in Conflict Terrorism	2010-2023	942	863
Territory Politics Governance	2013-2024	514	467
Terrorism and Political Violence	2010-2024	929	797
The Pacific Review	2010-2024	602	574
Third World Quarterly	2010-2024	1902	1722
Voluntas	2010-2024	1186	1150
West European Politics	2010-2024	1029	975
World Trade Review	2010-2024	477	355

F Calibration Inputs

Figure 13 shows how $p = (1 - \rho)$ varies with ϕ and π.

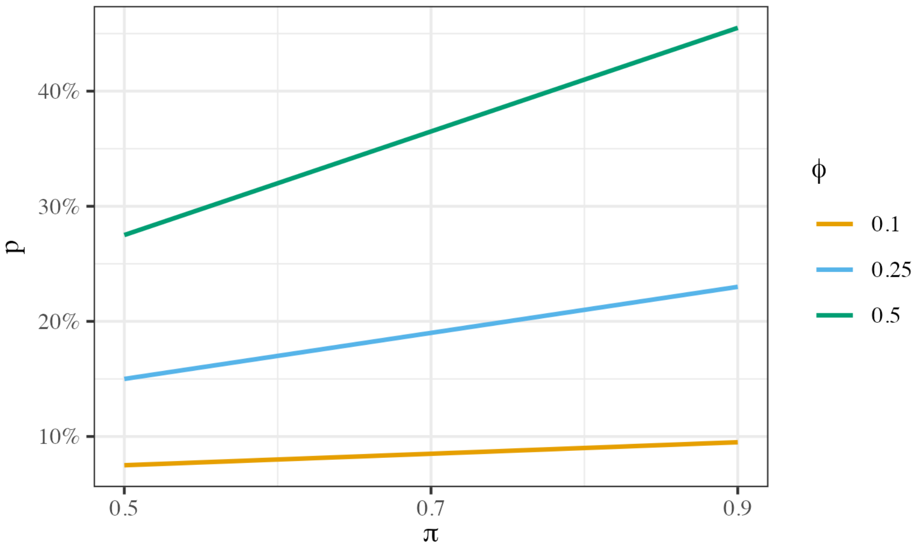

Figure 13: Statistical power (ϕ), share of true effects (π), and the single-test probability of significance ($1 - \rho$).