Can Large Language Models Role-Play Financial Regulators? Evidence from Climate Rulemaking*

Kirill Skobelev

Northwestern University

Rimmy E. Tomy

The Pennsylvania State University

November 19, 2026

Abstract

We examine whether large language models (LLMs) can role-play policymakers, and whether such role-playing can be used to uncover regulators' incentives. Our setting includes two climate rulemakings: the 2023 interagency guidance on climate-related financial risk management for large banks, and the SEC's 2022 proposed rule on climate-related disclosures. Our methodology consists of a two-stage procedure. In the first stage, for each of ten roles that represent different incentives, the LLM decides which public comments a regulator in a given role would accept. In the second stage, the LLM revises the draft guidance using only those accepted comments. The first-stage results are robust. Under every model we test, including a model fine-tuned on regulatory text, the *Republican* and *Top 1% Wealth* roles are less likely to accept a comment the more it favors climate action, and the *Democratic* and *Worldwide Welfare* roles are more likely to accept it, in both rulemakings. The second-stage results show that LLMs are unable to replicate the policies of human decision-makers by taking on role-conditioned personas. These results suggest that regulatory decision-making is complex and cannot be easily approximated by persona-conditioned LLMs.

*We thank seminar participants at the Accounting Summer Camp at the University of Padova and Temple University for helpful comments. This work is supported by funding from Penn State and the Center for Applied Artificial Intelligence (CAAI) at the University of Chicago Booth School of Business.

Email addresses: kirillskobelev2031@u.northwestern.edu (Kirill Skobelev), rtomy@psu.edu (Rimmy E. Tomy)

1. Introduction

The incentives driving regulators' decisions are rarely observable. Can building AI policymakers help uncover these hidden incentives? We investigate whether large language models (LLMs) can act as policymakers and whether such role-playing can reveal the incentives underlying regulatory decisions.

Extant research has used LLMs to simulate human decision-makers. When conditioned on a persona, LLMs have been found to reproduce the attitudes of specified subpopulations (Argyle et al., 2023; Park et al., 2024) and replicate established experimental findings (Aher et al., 2023; Cui et al., 2025). LLMs have also been treated as computational stand-ins for human subjects (Horton, 2023; Manning et al., 2024). However, several studies urge caution in using LLMs to simulate human decision-making. Persona-conditioned LLMs can match average opinions while understating their variation and shifting with the wording of the prompt (Bisbee et al., 2024). Hullman et al. (2026) argue that informal validation does not provide the statistical guarantees needed to treat LLM responses as substitutes for human responses. Gao et al. (2025) find that LLMs can fail to reproduce human behavior even in simple settings.

Recent studies extend the approach to institutional decision-makers, simulating members of the Federal Open Market Committee (Kazinnik & Sinclair, 2025), professional forecasters (Hansen et al., 2025), chief financial officers (Graham et al., 2026), and judges (Posner & Saran, 2026). Extending this line of research, we ask whether an LLM can act as a regulator and draft policy in a setting where the policy the regulator actually adopted is observable. In doing so, we ask whether persona-conditioned LLMs can uncover regulators' incentives.

Our setting consists of two climate rulemakings. The first produced the *Principles for Climate-Related Financial Risk Management for Large Financial Institutions*, supervisory guidance that the Board of Governors of the Federal Reserve System (Fed), the Federal Deposit Insurance Corporation (FDIC), and the Office of the Comptroller of the Currency (OCC) issued jointly in October 2023 (Office of the Comptroller of the Currency et al.,

2023), after each had published its own draft for comment. The agencies withdrew the guidance by October 2025. The second produced the SEC's rule on *The Enhancement and Standardization of Climate-Related Disclosures for Investors*, which the SEC proposed in March 2022 with an adoption date of March 2024 (SEC, 2022c, 2024b). However, the SEC stayed the rule pending judicial review, ended its defense of the rule in March 2025, and proposed rescinding it in May 2026. Both policies therefore went through the full notice-and-comment process before being abandoned. We study climate policy-making because it has the advantage that comment selections across different roles can more easily be validated. For example, we anticipate that a regulator with a liberal orientation would take a pro–climate position. This, in turn, enables us to confirm that the LLM is actually adopting the role it has been assigned.

Furthermore, in climate rulemaking, regulators' private incentives can shape the process, making it a particularly suitable context for identifying those incentives. The role of central banks and financial regulators in addressing climate-related financial risks is the subject of an ongoing debate (Brunnermeier & Landau, 2020; Campiglio et al., 2018; Cochrane, 2020; Hansen, 2022; Skinner, 2021). Proponents of a larger role argue that climate risks threaten financial and price stability, so that central banks cannot fulfill their mandates without addressing them (Bolton et al., 2020; Dikau & Volz, 2021). Opponents argue that central banks lack the democratic legitimacy to make environmental policy, and that doing so could cost them political support and, with it, their independence (Cochrane, 2020; Hansen, 2022). The debate has become increasingly politicized, in step with the partisan polarization of U.S. climate politics (Egan & Mullin, 2024).[1] Globally as well, Shears et al. (2025) find the extent to which central banks manage climate risks is associated with a country's climate politics rather than with its economic exposure to transition risks. Central banks' public

[1] See, for example, "Trump's Executive Orders Leave Imprint on the Fed," *The New York Times*, February 7, 2025; "Fed Quits Global Climate Risk Group Ahead of Trump Presidency," *The New York Times*, January 17, 2025; "Raskin's Bid for Fed Post Is Stalled," *The Wall Street Journal*, February 18, 2022; "Biden's Climate Promises Put Yellen in a Bind," *The Wall Street Journal*, December 13, 2021.

communication about climate is likewise driven mainly by institutional factors rather than by exposure to climate risk (Campiglio et al., 2025). A parallel debate surrounds securities regulation, where opponents of the SEC's climate disclosure rule questioned whether the agency had the authority to adopt it (Peirce, 2022). Despite this attention, little is known about how U.S. financial regulators weigh competing pressures when they write climate-related rules.

Our data consists of the draft and final climate policies, as well as comments submitted by the public on the draft policies. We begin with descriptive analyses.[2] Specifically, we use an LLM to score each comment on a five-point *Climate Engagement Score*, from 1 (strong opposition to climate action by the regulator) to 5 (strong advocacy of ambitious, binding climate measures), and validate a subset of the scores against the ratings of human raters recruited on Prolific. For comments received in response to the bank regulation, 74% favor climate action (score of 4 or 5). Private individuals and climate action groups strongly favor climate action, whereas banks oppose it. Comments that raise banks' fossil fuel investments are associated with higher scores, and comments about the effects on community banks with lower scores.

We match paragraphs between the draft and the final guidance and use an LLM to identify the substantive changes. We then relate the revisions to the policy to the comments by defining measures of alignment. For each comment and each paragraph of the final guidance, our alignment measure is an indicator that equals one if the final paragraph is closer to the comment than the draft paragraph it replaced. We measure closeness as the cosine similarity of text embeddings.[3] For the bank policy, a one-point increase in a comment's score is associated with a 2.5 percentage point decrease in the probability that the revision moved toward the comment. The revisions were thus less likely to move toward a comment the more strongly it advocated for climate action.

[2]We report these analyses, for both regulators, in the Online Appendix.

[3]Text embeddings convert text into numerical vectors that capture meaning, allowing us to measure how related different texts are.

The SEC comments are more evenly divided with close to 50% of all comments scoring 4 or 5, in favor of strong climate action. For these comments, private individuals lean against the rule, while climate and environmental groups and investors and asset managers support it. Law firms and trade associations oppose it the most. The topic most strongly associated with a low score is questioning the SEC's authority to require the disclosures. Finally, the relation between a comment's score and the direction of the revisions is again negative as for the banking rule, but weaker.

In our main analysis, we ask LLMs to act as regulators with different institutional and political incentives, and write the final policy based on the draft version of the policy and the comments it chooses to accept. We then compare each simulated revision with the policy the regulators actually adopted. We assign an LLM one of ten roles that capture institutional, political, and social pressures a financial regulator might face: the Fed's monetary policy mandate (*Monetary*); the bank regulators' mandate to keep the banking system safe and sound (*Banking*); a risk-averse bureaucrat concerned with career security (*Bureaucrat*); the economic interests of the average American (*American*); Democratic and Republican party preferences (*Democratic*, *Republican*); a regulator who wants to appear nonpartisan (*Nonpartisan*); the wealth of the top 1% of Americans (*Top 1% Wealth*); the welfare of the average person worldwide (*Worldwide Welfare*); and a baseline that asks the model to act in accordance with its own training (*Training*). For the SEC, the two mandate roles become *Capital Formation* and *Investor Protection*, two of the objectives in the SEC's mission, and the other eight roles are unchanged.

Our methodology follows two steps. In the first, the LLM decides, for each role and each comment, whether to accept any of the comment's proposals, and summarizes the comments it accepts. In the second step, the LLM revises the draft policy (the Fed's draft guidance, or the SEC's proposed rule) using only the proposals it accepted, while advancing the role's interests. We then compare each simulated revision with the policy the regulators actually adopted by measuring the overlap between the topics the two documents cover. We run this

procedure with several commercial LLMs and with an open-weight model that we fine-tune on final rules that U.S. financial regulators issued on other topics.

We find robust results in the first stage. Under every model we test (five commercial LLMs and the fine-tuned model, in each setting), the *Republican* and *Top 1% Wealth* roles are less likely to accept a comment the more it favors climate action, and the *Democratic* and *Worldwide Welfare* roles are more likely to accept it.

However, the second-stage results show that LLMs are unable to replicate the policies of human decision-makers by taking on role-conditioned personas. The role revision that comes closest to the adopted policy depends on the model that wrote the revisions. For the bank regulators, the Spearman correlations between the role rankings produced by GPT-5.2, the fine-tuned model, and Claude Sonnet 4.5 range from -0.47 to 0.41; for the SEC, they range from 0.05 to 0.33. For the bank regulators, for example, the *Training* role ranks first among the ten roles under GPT-5.2 but eighth and ninth under the other two models, and *Republican* ranks last under GPT-5.2 but second under Sonnet. No role is consistently closest to the adopted policy in either setting, and within each model the confidence intervals of most pairs of roles overlap.

We also test whether the models write the policies based on their memory rather than the prompts they were given. We find the revisions do not appear to draw on the models' memory of the final policies. The two commercial models that write them cannot reproduce the final texts when asked to. Although both models know the main decisions in the SEC's final rule, such as dropping the disclosure of Scope 3 emissions, every one of their revisions of the SEC's proposal still addresses Scope 3 emissions. LLM role-playing thus gives a robust signal about which comments a role would accept, but it cannot replicate the policy that the regulators adopted. These results suggest that regulatory decision-making is complex and cannot be easily approximated by persona-conditioned LLMs.

Our study makes two contributions. First, we contribute to the literature on using LLMs to simulate human decision-makers (Argyle et al., 2023; Park et al., 2024; Horton, 2023).

We test LLM role-playing on institutional decisions whose outcomes are observable, and document both the value and the limits of the approach. We find that role conditioning generally works at the comment-acceptance stage, where each decision is a single choice to accept or reject. Our results are consistent across models, and in directions we anticipated. However, persona-conditioned LLMs are unable to replicate the adopted policy at the policy-revision stage where the model writes the policy, a task that earlier studies have not, to our knowledge, evaluated against an adopted policy. There, the ranking of the roles changes with the model. This finding is consistent with prior work that LLMs' stated political positions change when they must answer in their own words rather than choose among fixed options (Röttger et al., 2024), and that the conclusions of LLM simulations depend on the choice of model and prompt (Cummins, 2026). LLM role-playing can therefore indicate which comments a regulator with given incentives would accept, but regulatory decision-making is complex and cannot be easily approximated by persona-conditioned LLMs. This conclusion adds to the evidence that LLMs should be used as surrogates for human responses with caution (Hullman et al., 2026; Gao et al., 2025).

Second, we contribute to the debate on the role of financial regulators in addressing climate change. Prior work analyzes what central banks say about climate change (Arseneau et al., 2022; Campiglio et al., 2025; Gupta, 2025) and the role they take in addressing it (Campiglio et al., 2018; Shears et al., 2025). We provide evidence on the rulemaking process itself, for both the bank regulators and the SEC, and move from what regulators say to how they respond to the public. In both settings, the revisions moved toward the positions of less climate-engaged commenters, and we find little indication that strong pro-climate advocacy shaped the final policy. In this we add to two literatures. Research on interest-group influence in notice-and-comment rulemaking finds that final rules tend to move toward business and well-resourced commenters (Yackee & Yackee, 2006; Carpenter et al., 2024), including at the Federal Reserve and the SEC (Libgober, 2020; Ban & You, 2019), and that firms use the comment process strategically (Bertrand et al., 2021). Research in accounting uses comment

letters to infer who gains from a rule (Hochberg et al., 2009). Two contemporaneous studies use LLMs to analyze the comment letters on the SEC's proposal: Lashitew & Mu (2025) score the sentiment of large firms' letters, and Pappas & Xu (2025) examine the sentiment, arguments, and influence of the letters. We differ from both in studying the bank regulators as well, validating the scores against human raters, measuring for each comment whether the revised paragraphs moved toward it, and asking whether LLM role-playing recovers the regulators' incentives.

2. Background

2.1. U.S. bank regulators and climate-related financial risk

Congress directs the Fed to promote maximum employment, stable prices, and moderate long-term interest rates. Although the Fed has no explicit financial stability mandate (Haltom & Weinberg, 2017), it has taken on an increasingly prominent role in this area. Its mission statement, publicized on its website, states that the Fed "promotes the stability of the financial system and seeks to minimize and contain systemic risks through active monitoring and engagement in the U.S. and abroad" (Board of Governors of the Federal Reserve System, 2025). Whether climate risk falls within these responsibilities is disputed.

Proponents of a larger role for central banks point to several channels through which climate risk could become financial risk (TCFD, 2017). Extreme weather events could impair borrowers and raise banks' nonperforming loans (physical risk). As economies shift to low-carbon technologies, banks exposed to carbon-intensive industries could suffer losses on stranded assets (transition risk). Climate change could also affect price stability through food and energy prices. Thus, proponents argue, addressing climate risk is not mission creep but a condition for price and financial stability (Bolton et al., 2020; Dikau & Volz, 2021).

On the other hand, opponents argue that by acting outside their explicit mandates, central banks could lose political support across administrations and so weaken their independence in monetary policy, and that they would put their credibility at risk by overstating

what is known about climate change (Cochrane, 2020; Hansen, 2022). A further concern is that supervisory pressure could discourage lending to carbon-intensive industries and distort the allocation of credit. Related evidence shows that climate considerations could influence credit allocation. Banks that make climate commitments lend less to high-emission firms, which do not reduce their emissions as a result (Kacperczyk & Peydró, 2026). Also, banks tighten credit terms for high-emission firms exposed to carbon pricing (Ivanov et al., 2024). In a similar vein, Skinner (2021) argues that the Fed has relatively limited legal authority to address climate change directly. The Fed itself has generally been reluctant to engage in climate policy. Chair Jerome Powell has said that the Fed is not, and will not be, a "climate policymaker," and that without explicit legislation it would be inappropriate to use monetary policy or supervisory tools to promote a greener economy. He has also said that supervisors should require banks to understand and manage their material risks, including the financial risks of climate change (Powell, 2023).[4]

In October 2023, the Fed, the FDIC, and the OCC jointly issued the *Principles for Climate-Related Financial Risk Management for Large Financial Institutions* (Office of the Comptroller of the Currency et al., 2023). The guidance set out a high-level framework in six areas: governance; policies, procedures, and limits; strategic planning; risk management; data, risk measurement, and reporting; and scenario analysis. It encouraged banks to integrate climate-related risks into traditional risk categories such as credit, liquidity, operational, and legal risk. It called for a risk-based approach tailored to each institution, and it made clear that an institution's public climate commitments should be consistent with its internal risk appetite and risk management framework. It also stressed that managing climate-related risks should not restrict lawful banking services or lead to discrimination against particular types of customers, especially low- and moderate-income (LMI) and other underserved consumers and communities.

Importantly, the guidance was not binding. It applied only to financial institutions with

[4]See also "Powell: Fed not seeking to become climate policymaker," *ABA Banking Journal*, April 3, 2024.

more than $100 billion in total consolidated assets, it set no specific requirements, and it was framed as a way to support large financial institutions to focus on key aspects of climate-related financial risk management. Even nonbinding guidance, however, can create implicit supervisory expectations that banks address climate risks, potentially including limits on lending to carbon-intensive industries. Banks that took part in the European Central Bank's 2022 climate risk stress test, for example, reduced their lending to emission-intensive industries even though the test had no capital consequences (De Cicco et al., 2025). On March 31, 2025, the OCC withdrew from the guidance, stating that it was "overly burdensome and duplicative" and that existing risk management frameworks adequately covered climate-related risks. The Fed and the FDIC withdrew from it in October 2025, and the three agencies published its rescission in the Federal Register in November 2025 (Office of the Comptroller of the Currency et al., 2025).[5]

Interestingly, bank regulators' engagement with climate policy has also followed the political cycle. The Network for Greening the Financial System (NGFS), formed in 2017, is an association of central banks and supervisors that seek to "exchange experiences, share best practices, contribute to the development of environment and climate risk management in the financial sector, and mobilize mainstream finance to support the transition toward a sustainable economy" (Network for Greening the Financial System, 2019). The U.S. bank regulators joined the NGFS between December 2020 and 2022, around the start of the Biden administration. The Fed became a member in December 2020, possibly in anticipation of the new administration's climate agenda, and later used NGFS climate scenarios in its pilot climate scenario analysis exercise (Federal Reserve, 2024; Acharya et al., 2023). The OCC joined in July 2021, when it also announced its first climate change risk officer, a position created to advance climate risk management practices at banks (OCC, 2021), and the FDIC joined in 2022. All three left as the Trump administration took office: the Fed announced

[5]See https://www.federalreserve.gov/newsevents/pressreleases/bcreg20251016a.htm (last accessed November 26, 2025).

its exit in January 2025, just before the inauguration, and the FDIC and the OCC followed in January and February 2025, respectively. These moves align with Republican opposition to climate-focused financial regulation. These events suggest that, in dealing with climate risk, the bank regulators are not purely technocratic institutions but actors navigating both economic and political pressures.

2.2. The SEC and climate-related disclosure

The SEC describes its mission as "protecting investors, maintaining fair, orderly, and efficient markets, and facilitating capital formation" (SEC, 2026a). In 2010, the SEC issued interpretive guidance explaining that existing requirements could oblige companies to discuss climate-related risks and opportunities in their descriptions of the business, legal proceedings, risk factors, and management's discussion and analysis (SEC, 2010). In March 2021, Acting Chair Allison Herren Lee requested public input on whether current disclosures adequately inform investors about climate change (Lee, 2021).

On March 21, 2022, the SEC proposed *The Enhancement and Standardization of Climate-Related Disclosures for Investors* (SEC, 2022c). The Commission approved the proposal by a 3–1 vote. The proposal would have required public companies to disclose climate-related risks that are reasonably likely to have a material impact on their business, their governance and management of those risks, climate-related financial metrics in their audited financial statements, and their greenhouse gas emissions, including Scope 3 emissions when material or when included in an emissions target. Commissioner Hester Peirce dissented, questioning whether the SEC had the authority to adopt such rules in a statement titled "We are Not the Securities and Environment Commission – At Least Not Yet" (Peirce, 2022). The rulemaking drew extensive public engagement. The SEC extended the comment period and later reopened it after a technological error in its comment system,[6] and it ultimately

[6]The original comment deadline of May 20, 2022, was extended to June 17, 2022 (SEC, 2022a). In October 2022, the SEC reopened the comment period until November 1, 2022, because a technological error had prevented it from receiving some comments (SEC, 2022b).

received more than 24,000 comment letters, including more than 4,500 unique letters (SEC, 2024a).

The proposal also fed a broader debate about mandatory climate disclosure. Mandated sustainability reporting can improve the information available to investors, but it imposes compliance costs, and its benefits are difficult to measure (Christensen et al., 2021). Studies report evidence consistent with the benefits of such mandates. Disclosure mandates around the world have increased stock liquidity (Krueger et al., 2024), and U.S. industrial facilities reduced their emissions after the Environmental Protection Agency began to publish them (Tomar, 2023). Institutional investors report that they value climate risk disclosures (Ilhan et al., 2023), but the evidence on how equity markets received the SEC's proposal is mixed. Shares of S&P 500 firms fell by 1.1% over the three days around its announcement, less so for firms with stronger environmental, social, and governance performance (Kim, 2024). Across the events leading up to and including the proposal, however, the average market reaction was not significant, and it was negative only for firms that did not already report their emissions (Frankel et al., 2022).

On March 6, 2024, the SEC adopted the final rules (SEC, 2024b). The final rules retained disclosure of material climate-related risks and of how firms govern and manage them, but they scaled back several proposed requirements. They dropped the disclosure of Scope 3 emissions, limited the disclosure of Scope 1 and Scope 2 emissions to larger registrants and to emissions that are material, and replaced the proposed financial statement metrics with narrower disclosure of the costs and losses from severe weather events and other natural conditions.[7] Challenges to the rules filed in several federal courts of appeals were consolidated in the Eighth Circuit, and on April 4, 2024, before the rules took effect, the SEC stayed them pending judicial review (SEC, 2024c).[8]

Following the change in administration, the SEC voted on March 27, 2025, to end its

[7] Online Appendix A documents these changes at the paragraph level.

[8] *Iowa v. SEC*, No. 24-1522 (8th Cir.). The Judicial Panel on Multidistrict Litigation consolidated the petitions for review on March 21, 2024.

defense of the rules, with Acting Chairman Mark T. Uyeda describing them as "costly and unnecessarily intrusive" (SEC, 2025). Rather than decide the petitions, the Eighth Circuit held the litigation in abeyance until the SEC either reconsidered the rules through notice-and-comment rulemaking or renewed its defense.[9] On May 29, 2026, the SEC proposed rescinding the rules in their entirety, arguing that they exceed its statutory authority and conflict with its objectives of facilitating capital formation and promoting public company status (SEC, 2026b). At the time of writing, the rescission has not been finalized, and the rules have never taken effect. Like the bank regulators' guidance, the SEC's rules therefore completed the full notice-and-comment process before the SEC abandoned them after the change in administration.

2.3. Two regulators as settings for LLM role-playing

The two rulemakings share the features that make them useful for testing whether LLMs can role-play policymakers. In each, regulators with explicit statutory mandates issued a draft, received public comments, and adopted a revised final policy. We therefore observe the comments the regulators received and the policy they chose, so we can ask which comments an LLM in a given role accepts and whether the policy it writes resembles the one the regulators adopted. Both rulemakings were also politically contested and were abandoned after the change in administration, which makes partisan and distributional roles plausible alternatives to roles grounded in each agency's mandate.

The settings also differ in ways that matter for this exercise. The bank regulators jointly issued nonbinding supervisory guidance, under mandates centered on price stability and the safety and soundness of the banking system. The SEC adopted binding disclosure rules, by commission votes that split along party lines, under a mission centered on investors and capital markets. We therefore tailor the two mandate-based roles to each regulator: *Monetary*

[9] Order, *Iowa v. SEC*, No. 24-1522 (8th Cir. Sept. 12, 2025). In a status report filed on July 23, 2025, the SEC had told the court that it did not intend to review or reconsider the rules at that time and had asked the court to decide the petitions (SEC, 2026b).

and *Banking* for the bank regulators, and *Capital Formation* and *Investor Protection*, two of the objectives in the SEC's mission, for the SEC. The SEC's rulemaking was also larger. Its rule text is almost four times as long as the Fed's draft guidance (about 7,500 words against 2,000), it drew about eight times as many comments, and its final rule departed further from the proposal.[10]

3. Data and sample

For each regulator, our data consist of the draft and final policies and the public comments on the drafts. Each of the three bank regulators issued its own draft of the guidance for comment. The OCC issued its draft on December 16, 2021, with comments due on February 14, 2022; the FDIC published its draft on April 4, 2022, with comments due on June 3, 2022; and the Fed issued its draft on December 2, 2022, with comments due on February 6, 2023. The final interagency guidance was published on October 30, 2023. We take the OCC's draft from the OCC's bulletin (Office of the Comptroller of the Currency, 2021), and the other two drafts (Federal Deposit Insurance Corporation, 2022; Board of Governors of the Federal Reserve System, 2022) and the final guidance (Office of the Comptroller of the Currency et al., 2023) from the Federal Register.[11]

We obtain the comments on the OCC's draft from *Regulations.gov*, the U.S. government's portal for public participation in federal rulemaking, and the comments on the FDIC's draft from the FDIC's website. The comments sent to the Fed are not publicly available, and we obtain them through a Freedom of Information Act (FOIA) request. Our sample consists of

[10] An LLM comparison of matched paragraphs flags as substantively different 67% of the pairs that match a paragraph of the SEC's final rule to its closest proposed paragraph, against 28% of the corresponding pairs for the bank regulators (Online Appendix A).

[11] The Federal Register is the official daily publication of the U.S. government for proposed rules, final rules, notices, and presidential documents. The drafts are at https://www.occ.gov/news-issuances/bulletins/2021/bulletin-2021-62a.pdf (OCC), https://www.govinfo.gov/content/pkg/FR-2022-04-04/pdf/2022-07065.pdf (FDIC), and https://www.govinfo.gov/content/pkg/FR-2022-12-08/pdf/2022-26648.pdf (Fed), and the final guidance is at https://www.govinfo.gov/content/pkg/FR-2023-10-30/pdf/2023-23844.pdf (all last accessed August 4, 2025).

581 comments: 314 sent to the OCC, 141 sent to the Fed, and 126 sent to the FDIC.[12]

The SEC issued its proposed rule (File No. S7-10-22) on March 21, 2022. The proposal was published in the Federal Register on April 11, 2022, with comments initially due on May 20, 2022. The SEC adopted the final rule on March 6, 2024, and published it in the Federal Register on March 28, 2024. We take both the proposed and the final rule from the Federal Register (SEC, 2022c, 2024b).[13] Unlike the bank regulators' guidance, both releases are long documents that combine the codified rule text with extensive discussion, economic analysis, and conforming amendments.

We collect the comments listed on the SEC's comment page for the file number and exclude 44 documents whose text exactly duplicates an earlier submission. Our sample consists of 4,892 documents: 4,465 individual comment letters, 394 memoranda of meetings between SEC officials and outside parties, and one copy of each of the 33 form letters that commenters submitted in large numbers, which together stand in for 20,078 submissions. The meeting memoranda matter because meetings with agency officials are a channel through which firms influence financial rules (Libgober, 2020; Ban & You, 2019). We count each form-letter campaign once so that every distinct text counts once, as for the bank regulators. Agencies also respond less to mass comment campaigns than to other comments (Balla et al., 2022). Our sample is therefore much smaller than the more than 24,000 comment letters that the SEC reports receiving, while its 4,465 letters are close to the more than 4,500 unique letters that the SEC reports (SEC, 2024a).[14]

[12]The FDIC's comments consist of 70 individual letters; 55 letters, mostly from community banks, that the Independent Community Bankers of America (ICBA) forwarded to the FDIC in a single submission and that we treat as separate comments; and a staff memorandum summarizing a meeting with several banks.

[13]The proposed rule is at `https://www.govinfo.gov/content/pkg/FR-2022-04-11/pdf/2022-06342.pdf` and the final rule at `https://www.govinfo.gov/content/pkg/FR-2024-03-28/pdf/2024-05137.pdf` (both last accessed September 25, 2026). Section 2.2 describes the extension and reopening of the comment period.

[14]Of the 4,892 documents, 1,633 are PDF files, which we convert to text with Marker, an open-source PDF parser (`https://github.com/datalab-to/marker`), and 3,259 are web-form submissions, which we convert with a standard HTML parser. For each of the 304 PDFs longer than ten pages, we classify every page as substantive or supporting material, and we truncate the 113 that contain supporting material, keeping the letter body and dropping appendices, exhibits, and signature pages. The longest submission, a signature petition, had 1,323 pages before truncation. In the processed text, the median document has 188 words and

4. Descriptive analyses

We run descriptive analyses of the policies and the comments for both regulators. We first document what changed between the draft and the final policy in each setting, by matching paragraphs across the two versions and using OpenAI's `gpt-4o-2024-08-06` to identify the substantive differences between matched paragraphs. We then use the same model to assign every comment a *Climate Engagement Score*, a measure from 1 to 5 of how strongly the comment favors climate action by the regulator. We validate a subset of the LLM-generated scores against the ratings of human raters, and find these to be consistent.[15]

We next turn to the content of the comments. We compare the score across types of commenter, extract the topics each comment raises, check that the topics are largely distinct, and regress the score on indicators for the topics. Among the comments to the bank regulators, climate action groups have a mean score of 4.91 and banks a mean of 2.01. Among the comments to the SEC, climate or environmental groups have a mean of 4.68, while trade or industry associations have a mean of 2.37 and issuers a mean of 2.86. Private individuals, who write most of the comments in both settings, differ sharply between the two: their mean score is 4.57 for the bank regulators but 2.93 for the SEC. Across all comments, the mean score is 4.03 for the bank regulators and 3.10 for the SEC. Furthermore, the score varies with the topics in the directions expected of a measure of support for climate action. For the bank regulators, comments about fossil fuel investments, a topic in 31% of the comments, score 1.018 points higher, holding the other topics fixed, whereas comments about community banks or regulatory burden score between 0.826 and 1.508 points lower. For the SEC, comments about climate impact disclosure, a topic in 4.6% of the documents, score 1.492 points higher, whereas comments that question the agency's

the 95th percentile has 4,906. Table OA5 in the Online Appendix summarizes the corpus.

[15] We report these analyses in the Online Appendix. Online Appendix A documents the policy changes, with examples in Table OA1 for the bank regulators and Table OA2 for the SEC. Online Appendix B describes the score. Its distribution appears in Figure OA1: for the bank regulators combined (Figure OA1(a)), for the SEC (Figure OA1(b)), and by agency (Figure OA1(c), Figure OA1(d), and Figure OA1(e)). Figure OA2 plots the mean model score against the human score, and Figure OA3 their joint distribution.

authority to require the disclosures score between 1.350 and 2.340 points lower. These results suggest that opponents of the bank guidance emphasize its burden, particularly on community banks, whereas opponents of the SEC rule contest the agency's authority to adopt it.[16]

Finally, we relate the policy changes to the comments. For every pair of a comment and a final policy paragraph, we construct an indicator that equals one when the final paragraph is closer to the comment than the draft paragraph it replaced, and we regress that indicator on the comment's score. We also examine a continuous version of the alignment measure by decile. In both settings, the comments that the final policy moved toward tend to come from less climate-engaged commenters.[17]

5. Role-playing regulators with LLMs

5.1. Constructing role-based policy proposals

We assign an LLM the role of a regulator with a given set of incentives and ask two questions: whether the LLM, in that role, behaves as such a regulator plausibly would, and whether the policies it writes reveal which incentives shaped the policy the regulators actually adopted. We describe the procedure for the bank regulators here and the adaptations for the SEC in Section 5.2. The LLM receives the draft policy, which for the bank regulators is the Fed's draft guidance, and the public comments, and it revises the draft in two stages. In the first stage, for each role and each comment, the LLM decides whether to accept or reject the comment's proposals based on the role, and it summarizes any proposed changes it accepts. In the second stage, the LLM receives the draft together with the summaries of all the proposals it accepted under the role, and it writes a single revised policy that

[16] Online Appendix C covers the content of the comments: Table OA3 and Table OA6 report the score by commenter type, Figure OA4 and Figure OA5 the correlations between topics, and Table OA4 and Table OA7 the estimates of Equation 2.

[17] Online Appendix D covers alignment: the indicator is defined in Equation 3 and the regression in Equation 4, with estimates in Table OA8 and Table OA9, and Figure OA6 plots the mean score by decile of the continuous measure for the bank regulators (Figure OA6(a)) and the SEC (Figure OA6(b)).

incorporates them while advancing the role's interests.

Our use of LLMs to simulate regulatory decision-making rests on the premise that these models can adopt a specific persona and act consistently with it. Recent research provides some support for this premise, while also highlighting important limitations. When conditioned on demographic or biographical information, LLMs reproduce the distribution of attitudes and responses observed in corresponding subpopulations (Argyle et al., 2023), and generative agents constructed from in-depth interviews replicate the attitudes of the individuals interviewed (Park et al., 2024). LLMs also reproduce classic findings in behavioral economics and social psychology (Aher et al., 2023) and the main effects of most scenario-based experiments in psychology and management, although with larger effect sizes than those found in the original studies (Cui et al., 2025). Consistent with this evidence, Horton (2023) describes LLMs as "homo silicus," computational models of human behavior whose responses to experimental scenarios can mirror established human findings.

At the same time, studies highlight the limitations of LLMs in reliably reproducing human decision-making. Hullman et al. (2026) caution that informal "validate-then-simulate" approaches, which treat simulated and human responses as interchangeable once they agree in a related setting, cannot rule out systematic biases in the simulated responses. Figure 1 summarizes the broader evidence by plotting the scale of each study's headline result against a five-point score capturing how strongly it supports the use of LLMs for role-playing. Eight of the seventeen studies provide supportive evidence. The strongest evidence (+2) comes from two large studies of persona conditioning: Argyle et al. (2023) use 2,873 human evaluators to compare texts written by people and by the model, while Park et al. (2024) construct 1,052 interview-based agents. Six additional studies provide support with modest caveats (+1). These include work that replicates economic experiments, simulates social scenarios, or applies LLMs to marketing research (Horton, 2023; Manning et al., 2024; Arora et al., 2025; Li et al., 2024), as well as studies that simulate institutional decision-makers such as CFOs and the FOMC committee (Graham et al., 2026; Kazinnik & Sinclair, 2025). Three

studies provide mixed evidence (0): LLMs replicate main effects but inflate effect sizes (Cui et al., 2025), exhibit a "hyper-accuracy distortion" in one of four experiments (Aher et al., 2023), or understate disagreement among professional forecasters (Hansen et al., 2025).

The remaining six studies challenge the reliability of LLM role-playing. Regression coefficients estimated from synthetic responses differ significantly from those estimated from 7,530 human survey respondents in nearly half of the cases (Bisbee et al., 2024); nearly all approaches fail to reproduce the distribution of human behavior across 8,000 model sessions (Gao et al., 2025); and an LLM deciding a case is influenced by precedent but not by the sympathetic portrayal of the defendant, contrary to the behavior of the federal judges in the original experiment (Posner & Saran, 2026). Other work argues that simulated responses should be statistically validated against, or combined with, human data (Hullman et al., 2026; Broska et al., 2025), and that seemingly defensible choices in designing a simulation can materially affect its conclusions (Cummins, 2026).

Taken together, this literature suggests that LLMs can capture meaningful features of how people in a given role behave, but that their simulated decisions may also contain systematic and context-dependent biases. Our study extends this literature to regulatory policymaking by using LLMs to simulate regulatory decision-making and examining the incentives that shape regulators' choices. Specifically, we compare policies generated by role-conditioned LLMs with the final policies adopted in practice, to test whether persona-conditioned LLMs can replicate regulators' decisions.

As described above, our methodology follows two stages. In the first stage, we provide the LLM with the following prompt:

> You will be given a comment submitted in response to a policy proposal. Your role is {role}. Your task is to review the comment and decide if you would accept any of the comment's proposals. The decision to accept or reject the proposal should be based on your role and nothing else. You may choose to accept parts of the comment or the comment in entirety. Act according to your role in full extent.
>
> Output two fields:

accepted: True if any part of the comment is accepted. False if all of it is ignored.
proposed_ change: brief summary of the accepted changes. Up to three sentences.

The second-stage prompt is as follows. The "public comments" it refers to are the summaries of the proposals accepted in the first stage.

> Your role is {role}, tasked with updating climate-related banking guidelines. You have access to public comments on the draft policy. Your objective is to revise the draft in line with your role's interests, using the comments solely to advance those interests. You may draw on logic, examples, arguments, and information from the comments, but only to strengthen your position. Some comments may be irrelevant or reflect opposing interests.
>
> You will be provided with:
>
> 1. The original draft of the policy.
> 2. A set of public comments submitted in response to that draft.
>
> You must output the revised version of the policy. The final version must be under 2000 words. The current policy is {words} words long. You must maintain the original paragraph order.

We ask the LLM to assume the following ten roles (short names in parentheses):

1. (*Monetary*) *Your role is*... to follow your mandate to maintain long run growth of the monetary and credit aggregates commensurate with the economy's long run potential to increase production, so as to promote effectively the goals of maximum employment, stable prices, and moderate long-term interest rates.

2. (*Banking*) *Your role is*... to ensure the safety and soundness of the banking system, and to make sure that the institutions you regulate and supervise understand and can effectively manage the material risks they face, including, climate-related financial risks.

3. (*Bureaucrat*) *Your role is*... that of a risk-averse bureaucrat who avoids risks that could threaten personal career security.

4. (*Democratic*) *Your role is*... to align policy with Democratic party preferences.

5. (*Republican*) *Your role is*... to align policy with Republican party preferences.

6. (*Nonpartisan*) *Your role is*... that of a regulator who wants to appear non-partisan.

7. (*American*) *Your role is*... to advance the economic interests of the average American.

8. (*Top 1% Wealth*) *Your role is*... to protect the wealth of the top 1 percent of Americans.

9. (*Worldwide Welfare*) *Your role is*... to promote the welfare of the average person worldwide.

10. (*Training*) *Your role is*... to act in accordance with the training you received from {developer}.

The roles capture incentives and pressures that financial regulators might face. Roles #1 and #2 are specific to the bank regulators; for the SEC, they are replaced by two objectives of its mission (Section 5.2). Role #1 follows the Fed's monetary policy mandate in Section 2A of the Federal Reserve Act, and role #2 the bank regulators' objective of keeping the banking system safe and sound. Role #3 is a risk-averse bureaucrat who avoids risks to their career security. Roles #4 and #5 represent partisan political pressures, aligning policy with Democratic or Republican party preferences, and role #6 represents a regulator who wants to appear nonpartisan. Role #7 advances the economic interests of the average American, whereas role #8 represents the possibility that regulatory decisions protect the wealth of the top 1% of Americans. Role #9 takes a global perspective, promoting the welfare of the average person worldwide. Finally, role #10 asks the LLM to act in accordance with the training it received from its own developer (OpenAI, Anthropic, or Google, depending on the model).

We run the full two-stage procedure ten times for each role, which gives ten revised policies per role and shows how much the output varies from run to run.[18] We also repeat

[18] We limit the procedure to ten runs because of the cost of each run.

the analysis with several LLMs, including commercial models and a model we fine-tune, to assess whether our findings depend on the choice of model.

A natural concern is that general-purpose commercial LLMs may not capture how U.S. financial regulators draft policy. To address it, we fine-tune an open-weight model and run the same two-stage procedure with it. Starting from Google's `gemma-3-27b-it` model (Gemma Team, 2025), we further train the model on a corpus of final rules issued by U.S. federal financial regulators, so that it learns the style of published regulatory text.[19] The fine-tuning thus exposes the model to how U.S. financial regulators write, without showing it the climate guidance whose revisions we study.[20]

5.2. Implementation for the SEC

We apply the same two-stage procedure to the SEC, with four adaptations. First, the two mandate roles are replaced by two objectives of the SEC's mission, capital formation and investor protection, and the other eight roles are unchanged:

1. (*Capital Formation*) *Your role is*... to facilitate capital formation and the efficient functioning of the securities markets, ensuring that any disclosure requirements are cost-effective and do not unduly burden registrants' access to public capital.

[19]The corpus consists of final rules issued by the OCC, the CFPB, the NCUA, and other federal financial regulators over roughly the past fifteen years, on routine topics such as bank capital, consumer protection, lending, and governance. None of the documents in the corpus concerns climate risk, and the 2022 climate-principles docket is deliberately excluded, so the model sees the regulators' general writing style but neither the draft it later revises nor the final guidance with which we compare its revisions. We convert the documents to markdown and split them into chunks of 2,048 tokens, which gives a training set of approximately 8 million tokens. We use LoRA (Low-Rank Adaptation; Hu et al., 2022), a fine-tuning method that freezes the original model's parameters and trains only a small set of additional parameters. This makes fine-tuning a 27-billion-parameter model feasible: we train LoRA adapters of rank 256 for three epochs on eight NVIDIA H100 GPUs.

[20]The fine-tuning corpus does not rule out that a model saw the policies during pretraining. The training data of Gemma 3 extend to August 2024 (`https://ai.google.dev/gemma/docs/core/model_card_3`), and those of the commercial models we use are at least as recent, so every model could in principle have seen the drafts and the final policies of both regulators, all of which were published by March 2024. Pretrained models can carry information about later outcomes into analyses meant to rely only on earlier data (Sarkar & Vafa, 2024; Lopez-Lira et al., 2025; Levy, 2026), which is why Ludwig et al. (2026) require that the researcher's sample not appear in a model's training data when LLM outputs are used for prediction. Such leakage could raise the similarity of every role's revisions to the final policy; it would bias the ranking of the roles only if it affected some roles more than others. In Section 5.7,we test whether the commercial models remember the final policies.

2. (*Investor Protection*) *Your role is*... to protect investors and ensure they receive consistent, comparable, and decision-useful climate-related disclosures so they can understand and price the material risks that registrants face.

Second, the policy that the LLM revises is the codified rule text of the proposal (Regulation S-X Article 14 and Regulation S-K Subpart 1500), which runs to about 7,500 words, against about 2,000 words for the Fed's draft guidance. The second-stage prompt is the same, except that it describes the task as "updating climate-related disclosure rules for public companies" and sets the word limit at 8,000 rather than 2,000. We compare each revision with the corresponding codified text of the final rule.

Third, because the SEC comment corpus is about eight times the size of the bank regulators', it changes how we run each stage. Each model makes its first-stage decisions once, and all ten runs reuse them, so run-to-run variation in the SEC results comes from the second stage only. GPT-5.2, our main model, and the fine-tuned Gemma evaluate all 4,892 comment documents. To limit cost, every other model evaluates the same random sample of 1,000 comment documents, stratified by deciles of comment length so that the sample matches the corpus in the distribution of comment length.

Fourth, on the full corpus the accepted proposals do not fit in a single second-stage prompt. We therefore use GPT-5.2 to condense every role's accepted proposals before the revision. The condensing prompt does not mention the role. It asks the model to merge proposals that request the same change, to keep every distinct proposal along with the identifiers of the comments that support it, and not to drop, weaken, strengthen, or add proposals. This reduces each role's list to between 580 and 1,434 distinct proposals (74,000 to 136,000 tokens), which we pass to the unchanged second-stage prompt. The role therefore enters only through the comments it accepted and through the revision itself, as it does for the bank regulators.[21] The 1,000-document sample needs no condensing. We condense the

[21] We condense in chunks of about 10,000 tokens and repeat the step on the merged list until a round shortens it by less than 10%, which takes two or three rounds. Any proposal whose comment identifier

fine-tuned Gemma's accepted proposals in the same way, again with GPT-5.2, because the lists of eight of its ten roles do not fit in its context window. The fine-tuned Gemma also differs in how it revises the proposal: it revises the codified text one section at a time. [22]

5.3. First stage: which comments each role accepts

We first examine which comments each role accepts, and whether the pattern is the same across models. For each model and role, we estimate the linear probability model

$$\text{Accepted}_{is} \;=\; \alpha \;+\; \beta \cdot \text{Climate Engagement Score}_{i} \;+\; \varepsilon_{is}, \tag{1}$$

where *Accepted* is an indicator for whether the model accepts comment i under the role in run s, and *Climate Engagement Score* is the comment's score, from 1 (strong opposition to climate action by the regulator) to 5 (strong advocacy). We estimate the model for six LLMs (GPT-5.2, GPT-5.5, Claude Sonnet 4.5, Claude Opus 4.7, Gemini 3 Flash, and the fine-tuned Gemma) and for the roles defined in Section 5.1. For GPT-5.2, Claude Sonnet 4.5, and the fine-tuned Gemma, we pool the ten runs, so that each comment enters once per run; GPT-5.5, Claude Opus 4.7, and Gemini 3 Flash have a single run. Because a model's decisions on the same comment are correlated across runs, we cluster standard errors by comment.[23]

Panel (a) of Figure 2 reports the estimates of β for the bank regulators. Their signs generally line up with the roles. The *Republican* and *Top 1% Wealth* roles have negative coefficients under every model, from -0.086 to -0.276, all significant at the 0.1% level. A

the model omits is carried forward unchanged, so no accepted proposal is lost. In a validation on the *Top 1% Wealth* role, whose full list fits within the input limit, revisions drawn from the full list and from the condensed list are equally similar to the final rule (embedding cosine similarity of 0.777 in both cases over three runs each, against a standard deviation of 0.009 across runs).

[22] For each role and run, the model revises each section with the unchanged second-stage prompt and the role's condensed list of accepted proposals, with the word limit scaled to the section's share of the codified text. We join the revised sections in their original order.

[23] Online Appendix B describes the construction and validation of the score. For the three models with a single run, each comment enters once, and clustering by comment is equivalent to using heteroskedasticity-robust standard errors. For the SEC, all ten runs reuse a single set of first-stage decisions (Section 5.2), so the pooled estimates and their clustered standard errors equal those from one run.

one-point increase in a comment's score lowers the probability that these roles accept it by 9 to 28 percentage points, depending on the model. The *Democratic* and *Worldwide Welfare* roles have positive coefficients under every model, from 0.046 to 0.237, also all significant at the 0.1% level. The mandate roles (*Monetary*, *Banking*) and the *Bureaucrat*, *Nonpartisan*, and *American* roles have positive coefficients that are smaller than the *Democratic* role's under every model, consistent with a modest pro-climate lean. The magnitudes vary considerably across models.

Panel (b) repeats the exercise for the SEC with the six models. The results for the partisan and distributional roles are consistent with those for the bank regulators. The *Republican* and *Top 1% Wealth* coefficients are negative and lie between -0.187 and -0.256 under the five commercial models and at -0.117 and -0.048 under the fine-tuned Gemma. The *Democratic* and *Worldwide Welfare* coefficients are positive and lie between 0.096 and 0.251, all significant at 0.1%.

Investor Protection role favors climate-engaged comments under every model (0.072 to 0.225), whereas *Capital Formation* leans against them under GPT-5.2, GPT-5.5, and Claude Opus 4.7 (-0.051 to -0.199), has coefficients indistinguishable from zero under Claude Sonnet 4.5 and Gemini 3 Flash, and leans slightly toward them under the fine-tuned Gemma (0.027). This differs from the bank regulators, whose two mandate roles both lean modestly toward climate-engaged comments.

In both settings, each role responds to a comment's position on climate action in the same direction under every model we test, with the exception of the SEC's *Capital Formation* role. The size of the response, however, varies across models. Overall, the first-stage results offer partial support for using LLMs to simulate how regulators respond to public comments on climate policy.

5.4. Second stage: the policies each role writes

We next examine the policies each role writes in the second stage. Each role produces ten revised policies, one per run. Because this stage is costly, we run it with three models:

GPT-5.2, the fine-tuned Gemma, and Claude Sonnet 4.5. To compare the revised policies with the final policy, we identify the topics each document covers. We measure the topic overlap between each simulated revised policy and the actual final policy to assess which role produces a draft that most closely resembles the policy ultimately adopted by the regulator. This comparison allows us to infer which considerations, and, by extension, which incentives, most strongly shape regulators' final policy decisions.

We first use OpenAI's `gpt-5` to extract the main topics of each policy document: the three agency drafts, the final guidance, and the revised policies from all ten runs. The prompt is:

> You will receive a policy document. Extract the main topics discussed in the document. Each topic should be a phrase of at most three words. Do not use hyphens or any special characters in topic names; use only letters and spaces. Return a list of topics.

We embed each topic phrase with OpenAI's `text-embedding-3-large` model and merge near-duplicates whose cosine similarity exceeds 0.90, which gives a vocabulary of 334 topics.

To summarize the topics, we group their embeddings into fifteen themes, repeatedly merging the two groups whose members are, on average, most similar to one another. An LLM labels each theme with two to four words; we give it all fifteen groups at once so that the labels are distinct. For each document and theme, we sum the TF-IDF weights of the theme's topics, which gives a document-by-theme matrix.[24]

Panel (a) of Figure 3 displays this matrix as a heatmap for the bank regulator. The first two rows are the Fed's draft and the final guidance. The other ten rows are the role-revised policies, averaged across the ten runs and ordered by the cosine similarity of their theme weights to those of the final guidance. The columns are ordered by total weight, from broad themes such as *Climate Risk Management* and *Governance and Disclosure* on the left to

[24]TF-IDF (term frequency–inverse document frequency) weights each topic by how distinctive it is to a document. Topics that appear in nearly every policy receive little weight, and topics that appear in only one or two documents receive more. The weights therefore emphasize what each role raises distinctively rather than what every policy says.

narrow ones such as *Stress Testing*, *Taxonomy Alignment*, and *Staff Training* on the right. For example, the revisions under the *Monetary* role place a weight of 14.9 on *Climate Risk Management*, the sum over the 124 topics in that theme, about 10% more than the final guidance (13.5).

Two patterns stand out. First, *Climate Risk Management* and *Governance and Disclosure* carry the most weight in every document, as one would expect of prudential guidance. The roles differ, however, in the types of risk they emphasize (as depicted in Figure OA7 of the Online Appendix). Second, the roles differ on the lower-weighted themes, with substantial variation across roles in the themes they emphasize.

Panel (b) of Figure 3 shows the corresponding matrix for the SEC. The topics and themes differ, and two themes, *Climate Risk Management* and *Disclosure and Attestation*, carry the most weight in the final rule and in the revisions of every role. The *Disclosure and Attestation* theme receives substantially greater weight in the final rule (48.5) than in the proposed rule (8.2), indicating that the SEC significantly expanded the prominence of this theme during the rulemaking process. Notably, revisions generated under every role also assign greater weight to this theme than does the original proposal. Because these role-specific revisions incorporate information from public comments, this common shift is consistent with the SEC responding to concerns raised during the comment process. Focusing solely on the *Disclosure and Attestation* theme, the *Republican* role produces the revision that most closely aligns with the final rule.

We next measure which role's revisions come closest to the final guidance. For each document, we form a vector whose j-th element is one if the document covers topic j and zero otherwise, and we measure the similarity of a revised policy to the final guidance by the cosine similarity of the two vectors (topic-frequency (TF) vectors). Because each element is binary, they record whether a topic appears rather than how often, and the cosine similarity of two such vectors equals the number of topics the two documents share divided by the geometric mean of the numbers of topics each covers.

Panel (a) of Figure 4 presents the results for the bank regulator. It reports the mean TF similarity between each role's revisions and the final guidance across the ten runs, with 95% confidence intervals. The means range from 0.49 (*Republican*) to 0.61 (*Training*). *Training*, *Nonpartisan*, and *Democratic* rank highest, and *American*, *Banking*, and *Republican* rank lowest. The intervals overlap substantially. Only the intervals of *Training* and *Nonpartisan* lie entirely above that of *Republican*, and every other pair of intervals overlaps.

Panel (b) of Figure 4 reports the TF similarities for the SEC. The means range from 0.33 (*Capital Formation*) to 0.43 (*Nonpartisan*), a narrower range than for the bank regulators, and the intervals again overlap substantially. Only that of *Nonpartisan* lies entirely above that of *Capital Formation*. *Nonpartisan* ranks first, followed by *Republican* and *Training*.

5.5. Replication with a fine-tuned model

The results so far come from a general-purpose commercial model. We now repeat the second-stage analysis with the model we fine-tuned on final rules issued by U.S. financial regulators (Section 5.1), which addresses the concern that a general-purpose model may not capture the regulators' drafting conventions.

We extract, merge, and cluster the topics of the fine-tuned Gemma's revisions as in Section 5.4, which gives a vocabulary of 264 topics, which we classify into themes. Figure 5 reports the themed heatmap. For the bank regulator, two themes, *Financial Risk Management* and *Governance and Compliance*, carry the most weight in the revisions of every role. Figure 6 reports the TF similarities under the fine-tuned Gemma. For the bank regulator, the means range from 0.55 (*Democratic*) to 0.63 (*Bureaucrat*). *Bureaucrat*, *Top 1% Wealth*, *American*, and *Worldwide Welfare* rank highest; *Republican*, *Nonpartisan*, *Monetary*, and *Training* come next; and *Banking* and *Democratic* rank lowest. Only the intervals of *Bureaucrat* and *Democratic* do not overlap. For the SEC, the TF similarities range from 0.33 (*Investor Protection*) to 0.44 (*Nonpartisan*), and only the interval of *Nonpartisan* lies entirely above those of *Investor Protection* and *Republican*.

Figure 7 plots each role's mean TF similarity under GPT-5.2 (horizontal axis) against

its mean under the fine-tuned Gemma (vertical axis). The two rankings bear little relation to each other: their Spearman rank correlation is −0.25. *Banking* ranks ninth under both models. *Training* and *Democratic*, however, which rank first and third under GPT-5.2, rank eighth and tenth under the fine-tuned Gemma, the largest moves of any role, while *Bureaucrat* moves the other way, from sixth under GPT-5.2 to first under Gemma.

5.6. Stability of the rankings across models

We repeat the second-stage analysis a third time, with Claude Sonnet 4.5.[25] Figure 8 plots each role's rank within each of the three models. The rankings differ sharply across models. For the Fed, *Bureaucrat* ranks first under the fine-tuned Gemma and Sonnet but sixth under GPT-5.2; *Democratic* ranks last under the fine-tuned Gemma and Sonnet but third under GPT-5.2; *Training* ranks first under GPT-5.2 but eighth and ninth under the other two models; and *Republican* ranks last under GPT-5.2 but second under Sonnet. The Spearman rank correlations between the three rankings are −0.25 (GPT-5.2 and the fine-tuned Gemma), −0.47 (GPT-5.2 and Sonnet), and 0.41 (the fine-tuned Gemma and Sonnet). No role is consistently closest to the final guidance.

For the SEC, the three rankings are only weakly related. The Spearman rank correlations are 0.12 between GPT-5.2 and Sonnet, 0.33 between GPT-5.2 and the fine-tuned Gemma, and 0.05 between the fine-tuned Gemma and Sonnet. *Nonpartisan* ranks first under GPT-5.2 and fine-tuned Gemma but eighth under Sonnet, and *Top 1% Wealth* ranks first under Sonnet but fourth under GPT-5.2 and second under the fine-tuned Gemma. *Training* ranks third under GPT-5.2 and Sonnet and fourth under the fine-tuned Gemma, but first under none.

Overall, while the first-stage results show role conditioning that holds across six LLMs in both settings (Figure 2), the second-stage results show that LLMs are unable to replicate

[25] Figure OA8 and Figure OA9 of the Online Appendix report the Sonnet results. As described in Section 5.2, Sonnet runs on the 1,000-document SEC sample, whereas GPT-5.2 and Gemma run on the full corpus.

the policies of human decision-makers by taking on role-conditioned personas. The revisions do not, on average, bring the drafts closer to the policies ultimately adopted. For the bank regulators, the Fed's original draft is as close to the final guidance as the closest GPT-5.2 role revision and closer than any revision produced by the fine-tuned Gemma. Similarly, for the SEC, the role revisions are, on average, about as close to the final rule as the original proposal (Figure 7, Panel A). Beyond these overall similarities, the revisions also fail to capture key substantive changes. For example, every GPT-5.2 and Claude Sonnet 4.5 revision of the SEC's proposal retains a provision on Scope 3 emissions, even though the SEC dropped that provision from the final rule (Figure 7, Panel B). The rankings of the roles differ across the three models, and within each model the confidence intervals of most pairs of roles overlap. These findings echo the caution of Hullman et al. (2026) about the "validate-then-simulate" strategy that researchers use to justify replacing human participants with LLMs.[26]

5.7. Do the models remember the final policies?

The models we use were trained on data that extend beyond the publication of both final policies. A model could therefore write a revision that resembles the final policy because it remembers that policy, rather than because of its role or the comments it accepted. We test whether the two commercial models that write the revisions, GPT-5.2 and Claude Sonnet 4.5, remember the final policies, in three ways. First, following Golchin & Surdeanu (2024), we give a model the first half of a paragraph of the final policy and ask it to write the next words, once telling it which document the paragraph comes from and asking it to reproduce the published text, and once asking it only to continue the text. A model that has memorized a text reproduces the continuation, and more so when it is told the source. Second, we give the model each draft paragraph that the final policy replaced, tell it the source, and ask it to write the paragraph as the regulators adopted it. Third, we ask the model four multiple-choice questions about the main decisions in the SEC's final rule, such as whether the rule

[26] Additional role-playing results appear in Online Appendix E.

requires the disclosure of Scope 3 emissions.

Neither model remembers the text of the final policies. Given the first half of a paragraph of the final guidance, neither model reproduces a single sequence of eight words from the rest of the paragraph. For the SEC's final rule, the models reproduce between 3.0% and 6.1% of the eight-word sequences that do not appear in the proposal. Telling GPT-5.2 the source raises this share from 3.0% to 5.9%, a difference at the margin of statistical significance, and telling Sonnet the source does not raise it. Asked to write the adopted version of each draft paragraph, both models return the draft almost unchanged: they write at most 0.9% of the four-word sequences that the regulators added, and they keep at least 92% of those that the regulators removed.

The models do, however, know the main decisions of the SEC, and both answer all four questions correctly. Their revisions, however, do not reflect this knowledge. Each of the 200 revisions of the SEC's proposal by the two models still addresses Scope 3 emissions, which the final rule does not mention, and none uses the final rule's threshold for disclosing the financial effects of severe weather events. A provision on the climate expertise of board members, which the final rule dropped, remains in 81 of GPT-5.2's revisions and 79 of Sonnet's. Also, whether the provision remains depends on the role: under Sonnet, the *Republican* role keeps it in one of ten runs and the *Democratic* role in all ten. The models' knowledge of the final policies is therefore unlikely to explain how close their revisions come to those policies.[27]

6. Conclusion

We study whether LLMs can role-play policymakers, using two climate rulemakings as our setting: the interagency guidance on climate-related financial risk management for large banks and the SEC's rule on climate-related disclosures. Notice-and-comment rulemaking is

[27] Online Appendix F describes the three tests and the check of the revisions, and Table OA10 reports the results.

particularly well suited to this analysis because we observe the regulator's initial proposal, the public comments it receives, and the policy it ultimately adopts. We can therefore give simulated regulators the same public input and compare the policies they produce with the policies adopted by the actual regulators.

We begin by examining how the actual regulators revised their proposals and how those revisions relate to the public comments. In both rulemakings, comments were polarized, and regulators were less likely to revise their proposals in the direction of comments that more strongly advocated climate action. Thus, despite substantial public engagement, we find little evidence that highly climate-engaged commenters had greater influence on the final policies.

We then use LLMs to simulate regulators with different incentives. The roles include regulators guided by each agency's institutional mandate, a risk-averse bureaucrat, partisan political actors, and an advocate for global welfare. The simulation proceeds in two stages. First, for each public comment, the LLM determines whether a regulator in a given role would accept any of its proposals. Second, the LLM revises the proposed policy using the accepted proposals while pursuing the priorities associated with that role. We repeat this exercise across several commercial LLMs and an open-weight model fine-tuned on final rules issued by regulators in other policy domains.

The results differ sharply across the two stages. At the comment-acceptance stage, the patterns are remarkably robust across the five commercial LLMs and the fine-tuned model, and across both rulemakings. Under every model, comments that more strongly favor climate action are less likely to be accepted by the *Republican* and *Top 1% Wealth* roles and more likely to be accepted by the *Democratic* and *Worldwide Welfare* roles. At the policy-revision stage, however, LLMs are unable to replicate the policies that the regulators adopted. Which role produces a policy most similar to the regulator's final policy depends heavily on the LLM used to generate the revision. Indeed, the rankings of roles produced by different models are nearly unrelated, and no role consistently generates the policy closest to the one ultimately

adopted in either rulemaking.

These findings reinforce cautions in the literature about using LLMs as surrogates for human decision-makers (Gao et al., 2025; Hullman et al., 2026). LLM role-playing appears most reliable when models make many relatively simple decisions whose aggregate patterns can be compared across models, as in our comment-acceptance stage. It is considerably less reliable when models must synthesize those decisions into a complex policy document. This distinction is consequential: the policy-writing stage is precisely the stage needed to infer which simulated incentives best reproduce regulators' observed choices. When the resulting policy depends strongly on the choice of LLM, similarity between a simulated and actual policy provides a fragile basis for drawing conclusions about regulators' underlying incentives.

Future work could investigate whether additional calibration makes this policy-generation stage more reliable. For example, simulated policies could be calibrated against human judgments, as proposed by Hullman et al. (2026), or combined with human responses through mixed-subject designs (Broska et al., 2025). More broadly, our results suggest that LLMs can consistently adopt distinct policymaker roles and evaluate public comments in ways that reflect those roles' incentives. This consistency breaks down, however, when LLMs must translate those judgments into complex policy. Regulatory decision-making is thus complex and cannot be easily approximated by persona-conditioned LLMs.

References

Acharya, V. V., Berner, R., Engle, R., Jung, H., Stroebel, J., Zeng, X., & Zhao, Y. (2023). Climate stress testing. *Annual Review of Financial Economics*, *15*, 291–326.

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 337–371). PMLR volume 202 of *Proceedings of Machine Learning Research*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*, 337–351.

Arora, N., Chakraborty, I., & Nishimura, Y. (2025). AI–human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators. *Journal of Marketing*, *89*, 43–70.

Arseneau, D. M., Drexler, A., & Osada, M. (2022). Central bank communication about climate change. Working Paper.

Balla, S. J., Beck, A. R., Meehan, E., & Prasad, A. (2022). Lost in the flood? Agency responsiveness to mass comment campaigns in administrative rulemaking. *Regulation & Governance*, *16*, 293–308.

Ban, P., & You, H. Y. (2019). Presence and influence in lobbying: Evidence from Dodd-Frank. *Business and Politics*, *21*, 267–295.

Bertrand, M., Bombardini, M., Fisman, R., Hackinen, B., & Trebbi, F. (2021). Hall of mirrors: Corporate philanthropy and strategic advocacy. *Quarterly Journal of Economics*, *136*, 2413–2465.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, *32*, 401–416.

Board of Governors of the Federal Reserve System (2022). Principles for Climate-Related Financial Risk Management for Large Financial Institutions. Request for comment, 87 FR 75267–75271. URL: https://www.federalregister.gov/d/2022-26648.

Board of Governors of the Federal Reserve System (2025). About the Fed. https://www.federalreserve.gov/aboutthefed.htm. Accessed August 29, 2025.

Bolton, P., Despres, M., Da Silva, L. A. P., Samama, F., & Svartzman, R. (2020). *The Green Swan: Central Banking and Financial Stability in the Age of Climate Change*. Bank for International Settlements.

Broska, D., Howes, M., & van Loon, A. (2025). The mixed subjects design: Treating large language models as (potentially) informative observations. *Sociological Methods & Research*, *54*, 1074–1109.

Brunnermeier, M., & Landau, J.-P. (2020). Central banks and climate change. VoxEU. URL: https://voxeu.org/article/central-banks-and-climate-change.

Campiglio, E., Dafermos, Y., Monnin, P., Ryan-Collins, J., Schotten, G., & Tanaka, M. (2018). Climate change challenges for central banks and financial regulators. *Nature climate change*, *8*, 462–468.

Campiglio, E., Deyris, J., Romelli, D., & Scalisi, G. (2025). Warning words in a warming world: Central bank communication and climate change. *European Economic Review*, *178*, 105101.

Carpenter, D., Dagonel, A., Judge-Lord, D., Kenny, C. T., Libgober, B., Rashin, S., Waggoner, J., & Yackee, S. W. (2024). Inequality in administrative democracy: Large-sample evidence from American financial regulation. Working paper, SSRN 4770520.

Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: economic analysis and literature review. *Review of Accounting Studies*, *26*, 1176–1248.

Cochrane, J. H. (2020). Challenges for central banks. ECB Conference on Monetary Policy. URL: https://www.ecb.europa.eu/press/conferences/shared/pdf/20201019_confmonpol/Cochrane.pdf.

Costa, M., Desmarais, B. A., & Hird, J. A. (2019). Public comments' influence on science use in U.S. rulemaking: The case of EPA's national emission standards. *American Review of Public Administration*, *49*, 36–50.

Cui, Z., Li, N., & Zhou, H. (2025). A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, *5*, 627–634.

Cummins, J. (2026). The threat of analytic flexibility in using large language models to simulate human data. *Advances in Methods and Practices in Psychological Science*, *9*.

De Cicco, V., Gschossmann, I., & Kok, C. (2025). Bank lending implications of climate stress tests. Working paper, SSRN 5382449.

Dikau, S., & Volz, U. (2021). Central bank mandates, sustainability objectives and the promotion of green finance. *Ecological Economics*, *184*, 107022.

Egan, P. J., & Mullin, M. (2024). US partisan polarization on climate change: Can stalemate give way to opportunity? *PS: Political Science & Politics*, *57*, 30–35.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. URL: https://arxiv.org/abs/1909.00512.

Federal Deposit Insurance Corporation (2022). Statement of principles for climate-related financial risk management for large financial institutions; notice of proposed policy statement; request for comment (rin 3064–za32). Federal Register, 87(64), 19507–19512. URL: https://www.fdic.gov/resources/regulations/federal-register-publications/index.html notice of proposed policy statement.

Federal Reserve (2024). *Pilot Climate Scenario Analysis Exercise: Summary of Participants' Risk-Management Practices and Estimates*. Technical Report Board of Governors of the Federal Reserve System Washington, DC. URL: https://www.federalreserve.gov/publications/files/csa-exercise-summary-20240509.pdf accessed August 17, 2025.

Frankel, R. M., Lin, Y., & Martin, X. (2022). Does market demand for climate-related disclosure regulation? Working paper, SSRN 4278875.

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, *122*, e2501660122.

Gemma Team (2025). Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Golchin, S., & Surdeanu, M. (2024). Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations*.

Graham, J. R., Harvey, C. R., & Jha, M. (2026). CFOs meet LLMs. Working paper, arXiv:2606.13812.

Gupta, B. (2025). Talking the talk vs. walking the walk: greening of central banks' speeches and policies. *Climate and Development*, (pp. 1–20).

Haltom, R., & Weinberg, J. A. (2017). *Does the Fed Have a Financial Stability Mandate?*. Economic Brief (EB17-06) Federal Reserve Bank of Richmond. URL: https://www.richmondfed.org/-/media/richmondfedorg/publications/research/economic_brief/2017/pdf/eb_17-06.pdf.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., & Zarifhonarvar, A. (2025). Simulating the survey of professional forecasters. Working paper, SSRN 5066286.

Hansen, L. P. (2022). Central banking challenges posed by uncertain climate change and natural disasters. *Journal of Monetary Economics*, *125*, 1–15.

Hochberg, Y. V., Sapienza, P., & Vissing-Jørgensen, A. (2009). A lobbying approach to evaluating the Sarbanes-Oxley Act of 2002. *Journal of Accounting Research*, *47*, 519–583.

Horton, J. J. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?*. Technical Report National Bureau of Economic Research.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hullman, J., Broska, D., Sun, H., & Shaw, A. (2026). This human study did not involve human subjects: Validating llm simulations as behavioral evidence. Working Paper.

Ilhan, E., Krueger, P., Sautner, Z., & Starks, L. T. (2023). Climate risk disclosure and institutional investors. *Review of Financial Studies*, *36*, 2617–2650.

Ivanov, I. T., Kruttli, M. S., & Watugala, S. W. (2024). Banking on carbon: Corporate lending and cap-and-trade policy. *Review of Financial Studies*, *37*, 1640–1684.

Kacperczyk, M., & Peydró, J.-L. (2026). Carbon emissions and the bank-lending channel. Forthcoming in *Review of Financial Studies*.

Kazinnik, S., & Sinclair, T. M. (2025). FOMC in silico: A multi-agent system for monetary policy decision modeling. Working Paper 2025-005, Center for Economic Research, The George Washington University; SSRN 5424097.

Kim, M. M. (2024). The market reaction of S&P 500 firms to the SEC's mandatory climate disclosure proposal. *Journal of Corporate Accounting & Finance*, *35*, 110–120.

Klüver, H. (2009). Measuring interest group influence using quantitative text analysis. *European Union Politics*, *10*, 535–549.

Krueger, P., Sautner, Z., Tang, D. Y., & Zhong, R. (2024). The effects of mandatory ESG disclosure around the world. *Journal of Accounting Research*, *62*, 1795–1847.

Lagakos, D., Michalopoulos, S., & Voth, H. (2025). American life histories: What does it take to live a meaningful life? NBER Working Paper.

Lashitew, A., & Mu, Y. (2025). Corporate opposition to climate change disclosure regulation in the United States. *Climate Policy*, *25*, 422–437.

Lee, A. H. (2021). Public Input Welcomed on Climate Change Disclosures. Statement, U.S. Securities and Exchange Commission. URL: https://www.sec.gov/newsroom/speeches-statements/lee-climate-change-disclosures.

Levy, B. (2026). Caution ahead: Numerical reasoning and look-ahead bias in AI models. *Journal of Accounting Research*, *64*, 1139–1188.

Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, *43*, 254–266.

Libgober, B. (2020). Meetings, comments, and the distributive politics of rulemaking. *Quarterly Journal of Political Science*, *15*, 449–481.

Linder, F., Desmarais, B., Burgess, M., & Giraudy, E. (2020). Text as policy: Measuring policy similarity through bill text reuse. *Policy Studies Journal*, *48*, 546–574.

Lopez-Lira, A., Tang, Y., & Zhu, M. (2025). The memorization problem: Can we trust LLMs' economic forecasts? Working paper, arXiv:2504.14765.

Ludwig, J., Mullainathan, S., & Rambachan, A. (2026). Large language models: An applied econometric framework. *Annual Review of Economics*, *18*, 283–316.

Manning, B. S., Zhu, K., & Horton, J. J. (2024). *Automated social science: Language models as scientist and subjects*. Technical Report National Bureau of Economic Research.

Network for Greening the Financial System (2019). *A Call for Action: Climate Change as a Source of Financial Risk*. First Comprehensive Report Network for Greening the Financial System. URL: https://www.ngfs.net/en/first-comprehensive-report-call-action.

OCC (2021). OCC announces climate change risk officer, membership in the NGFS. News Release 2021-78. URL: https://www.occ.gov/news-issuances/news-releases/2021/nr-occ-2021-78.html accessed August 17, 2025.

Office of the Comptroller of the Currency (2021). Principles for Climate-Related Financial Risk Management for Large Banks. OCC Bulletin 2021-62, request for feedback. URL: https://www.occ.gov/news-issuances/bulletins/2021/bulletin-2021-62a.pdf.

Office of the Comptroller of the Currency, Board of Governors of the Federal Reserve System, & Federal Deposit Insurance Corporation (2023). Principles for Climate-Related Financial Risk Management for Large Financial Institutions. Final interagency guidance, 88 FR 74183–74189. URL: https://www.federalregister.gov/d/2023-23844.

Office of the Comptroller of the Currency, Board of Governors of the Federal Reserve System, & Federal Deposit Insurance Corporation (2025). Rescission of Principles for Climate-Related Financial Risk Management for Large Financial Institutions. Notice, 90 FR 51756–51757. URL: https://www.federalregister.gov/d/2025-20213.

Pappas, K., & Xu, A. L. (2025). From comments to policy: AI-assisted analysis on stakeholder responses to the SEC's climate disclosure proposal. Working paper, SSRN 5366251.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. Working Paper.

Peirce, H. M. (2022). We are Not the Securities and Environment Commission – At Least Not Yet. Statement, U.S. Securities and Exchange Commission. URL: https://www.sec.gov/newsroom/speeches-statements/peirce-climate-disclosure-20220321.

Posner, E. A., & Saran, S. (2026). Judge AI: A case-study of large language models as judges. *Journal of Law and Empirical Analysis*, *3*, 179–203.

Powell, J. H. (2023). Panel on "Central Bank Independence and the Mandate—Evolving Views". Speech at the Symposium on Central Bank Independence, Sveriges Riksbank, Stockholm, January 10. URL: https://www.federalreserve.gov/newsevents/speech/powell20230110a.htm.

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15295–15311).

Sarkar, S. K., & Vafa, K. (2024). Lookahead bias in pretrained language models. Working paper, SSRN 4754678.

SEC (2010). Commission Guidance Regarding Disclosure Related to Climate Change. Release Nos. 33-9106; 34-61469, 75 FR 6290. URL: https://www.sec.gov/files/rules/interp/2010/33-9106.pdf.

SEC (2022a). SEC extends comment period for proposed rules on climate-related disclosures, reopens comment periods for proposed rules regarding private fund advisers and regulation ATS. Press Release 2022-82. URL: https://www.sec.gov/newsroom/press-releases/2022-82.

SEC (2022b). SEC reopens comment periods for several rulemaking releases due to technological error in receiving certain comments. Press Release 2022-186. URL: https://www.sec.gov/newsroom/press-releases/2022-186.

SEC (2022c). The Enhancement and Standardization of Climate-Related Disclosures for Investors. Proposed rule, Release Nos. 33-11042; 34-94478, 87 FR 21334. URL: https://www.federalregister.gov/d/2022-06342.

SEC (2023). Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure. Final rule, Release No. 33-11216, 88 FR 51896. URL: https://www.federalregister.gov/d/2023-16194.

SEC (2024a). SEC adopts rules to enhance and standardize climate-related disclosures for investors. Press Release 2024-31. URL: https://www.sec.gov/newsroom/press-releases/2024-31.

SEC (2024b). The Enhancement and Standardization of Climate-Related Disclosures for Investors. Final rule, Release No. 33-11275, 89 FR 21668. URL: https://www.federalregister.gov/d/2024-05137.

SEC (2024c). The Enhancement and Standardization of Climate-Related Disclosures for Investors; Delay of Effective Date. Release No. 33-11280, 89 FR 25804. URL: https://www.federalregister.gov/d/2024-07648.

SEC (2025). SEC votes to end defense of climate disclosure rules. Press Release 2025-58. URL: https://www.sec.gov/newsroom/press-releases/2025-58.

SEC (2026a). Mission. https://www.sec.gov/about/mission. Accessed September 14, 2026.

SEC (2026b). Rescission of Climate-Related Disclosure Rules. Proposed withdrawal of final rules, Release Nos. 33-11421; 34-105572. URL: https://www.sec.gov/files/rules/proposed/2026/33-11421.pdf.

Shears, E., Meckling, J., & Finnegan, J. J. (2025). How central banks manage climate and energy transition risks. *Nature Energy*, (pp. 1–9).

Skinner, C. P. (2021). Central banks and climate change. *Vanderbilt Law Review*, *74*, 1301–1364.

TCFD (2017). Final report: Recommendations of the task force on climate-related financial disclosures. URL: https://assets.bbhub.io/company/sites/60/2020/10/FINAL-2017-TCFD-Report-11052018.pdf.

Tomar, S. (2023). Greenhouse gas disclosure and emissions benchmarking. *Journal of Accounting Research*, *61*, 451–492.

Yackee, J. W., & Yackee, S. W. (2006). A bias towards business? Assessing interest group influence on the U.S. bureaucracy. *Journal of Politics*, *68*, 128–139.

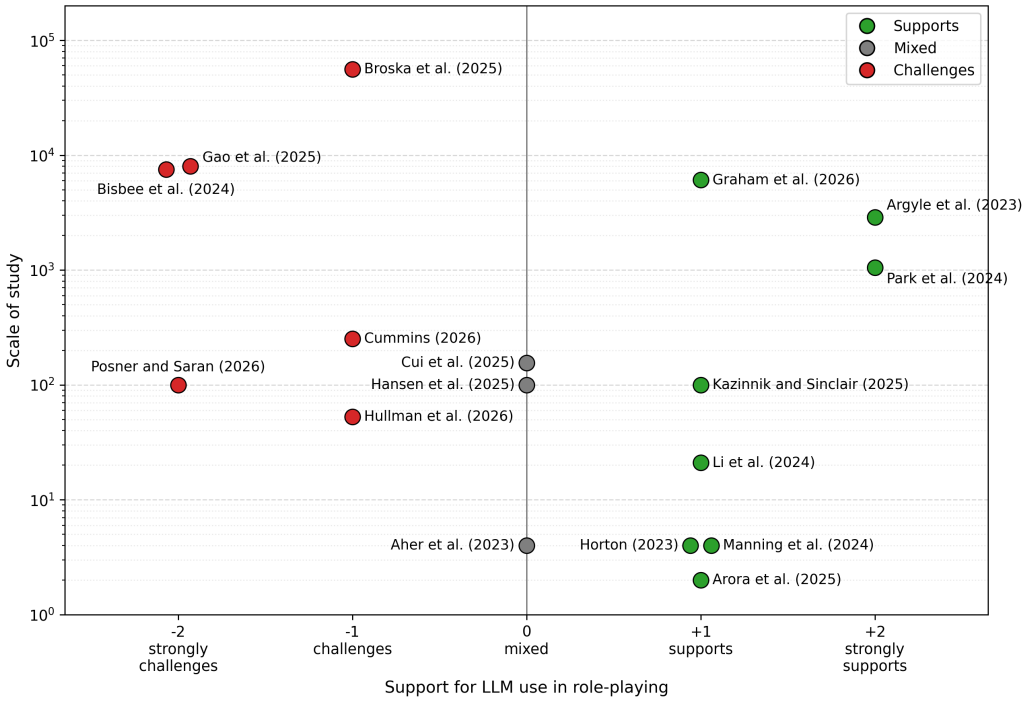


Figure 1: LLM role-playing literature

Each point is a paper cited in this study that examines whether LLMs can adopt personas, simulate specific human types, or replicate human behavior in experiments or surveys. The x-axis is a five-point supportiveness score assigned by an LLM after reading each paper's abstract and introduction: +2 strongly supports (LLMs reliably reproduce human responses on the tested dimension), +1 supports with modest caveats, 0 mixed (replicates main effects but documents systematic distortions such as hyperaccuracy or effect-size inflation), −1 challenges (LLM simulations require statistical rather than heuristic validation), −2 strongly challenges (documents non-random failures that distort substantive conclusions). The y-axis (log scale) is the headline study-scale number reported in each paper. Units differ across papers (human evaluators, survey respondents, simulated individuals or committee meetings, model decisions or sessions, experiments or company studies replicated, brands rated, simulation configurations, or studies drawn on), so the axis captures the empirical scope of each exercise rather than a directly comparable sample size. Points that would coincide, Horton (2023) and Manning et al. (2024), and nearly coincide, Bisbee et al. (2024) and Gao et al. (2025), are shifted slightly apart horizontally. Green markers denote papers whose findings primarily support LLM role-playing, gray markers denote mixed findings, and red markers denote findings that primarily challenge LLM role-playing.

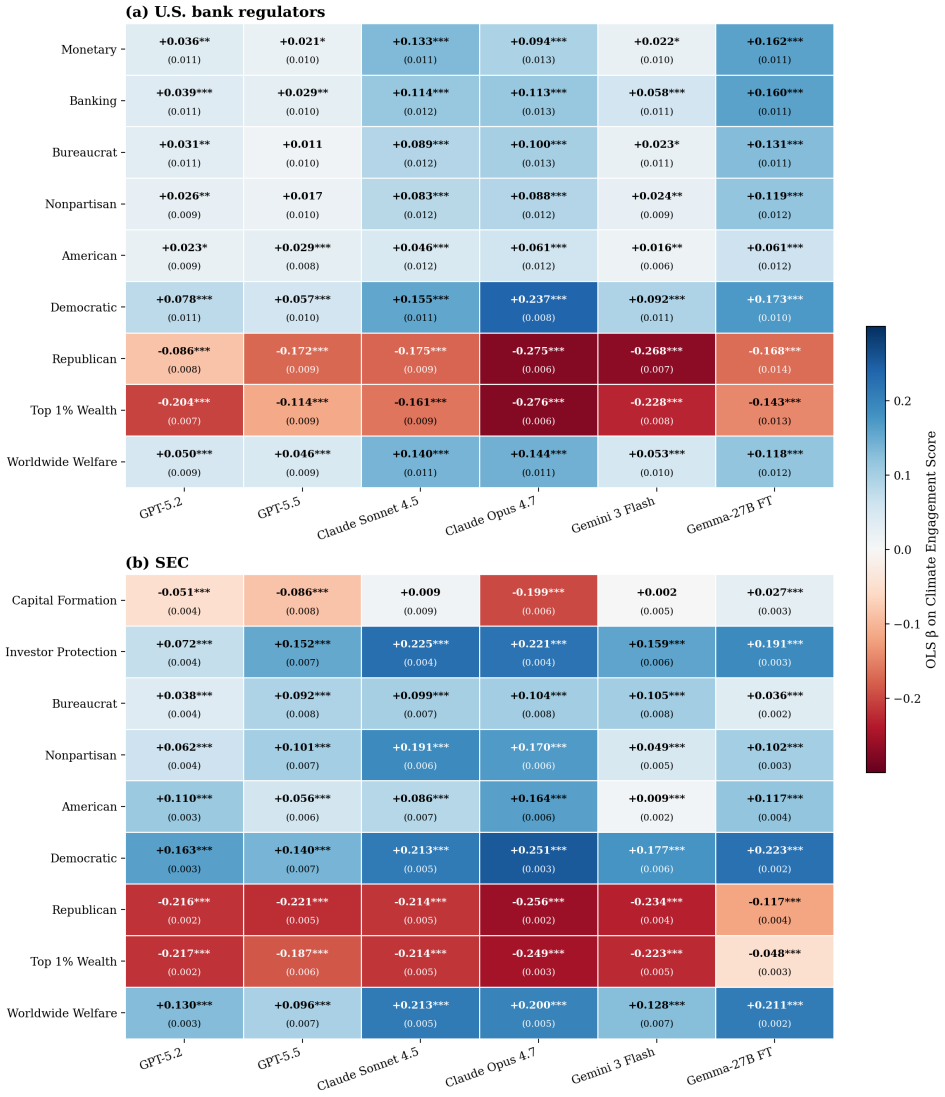


Figure 2: Role conditioning of comment acceptance: OLS coefficient of *Accepted* on *Climate Engagement Score*, by model and role, for (a) U.S. bank regulators and (b) the SEC.

Each cell reports the OLS slope coefficient β from the regression $Accepted_{is} = \alpha + \beta \cdot Climate\ Engagement\ Score_i + \varepsilon_{is}$ estimated separately for each model-role combination, where $Accepted_{is}$ is an indicator that equals one if the model accepts comment i in the first stage under the assigned role in run s, and $Climate\ Engagement\ Score_i$ is the comment's score, from 1 (strong opposition to climate action) to 5 (strong climate advocacy). The regressions pool the ten runs of GPT-5.2, Claude Sonnet 4.5, and the fine-tuned Gemma; the other three models have a single run. Standard errors are clustered by comment and reported in parentheses. Stars denote significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Columns span six LLMs: GPT-5.2, GPT-5.5, Claude Sonnet 4.5, Claude Opus 4.7, Gemini 3 Flash, and the fine-tuned Gemma (labeled Gemma-27B FT). Blue cells indicate roles that are more likely to accept a comment the higher its score, and red cells roles that are less likely to accept it. Panel (a) uses the 581 comments on the interagency climate guidance for large banks; panel (b) uses a random sample of 1,000 of the 4,892 comment documents on the SEC climate-related disclosure rule, stratified by comment length, except for GPT-5.2 and the fine-tuned Gemma, which use the full SEC corpus. Rows are aligned across panels: the SEC mandate roles *Capital Formation* and *Investor Protection* take the place of the bank-regulator mandate roles *Monetary* and *Banking*, and the remaining seven roles use identical prompts. Both panels share one color scale.

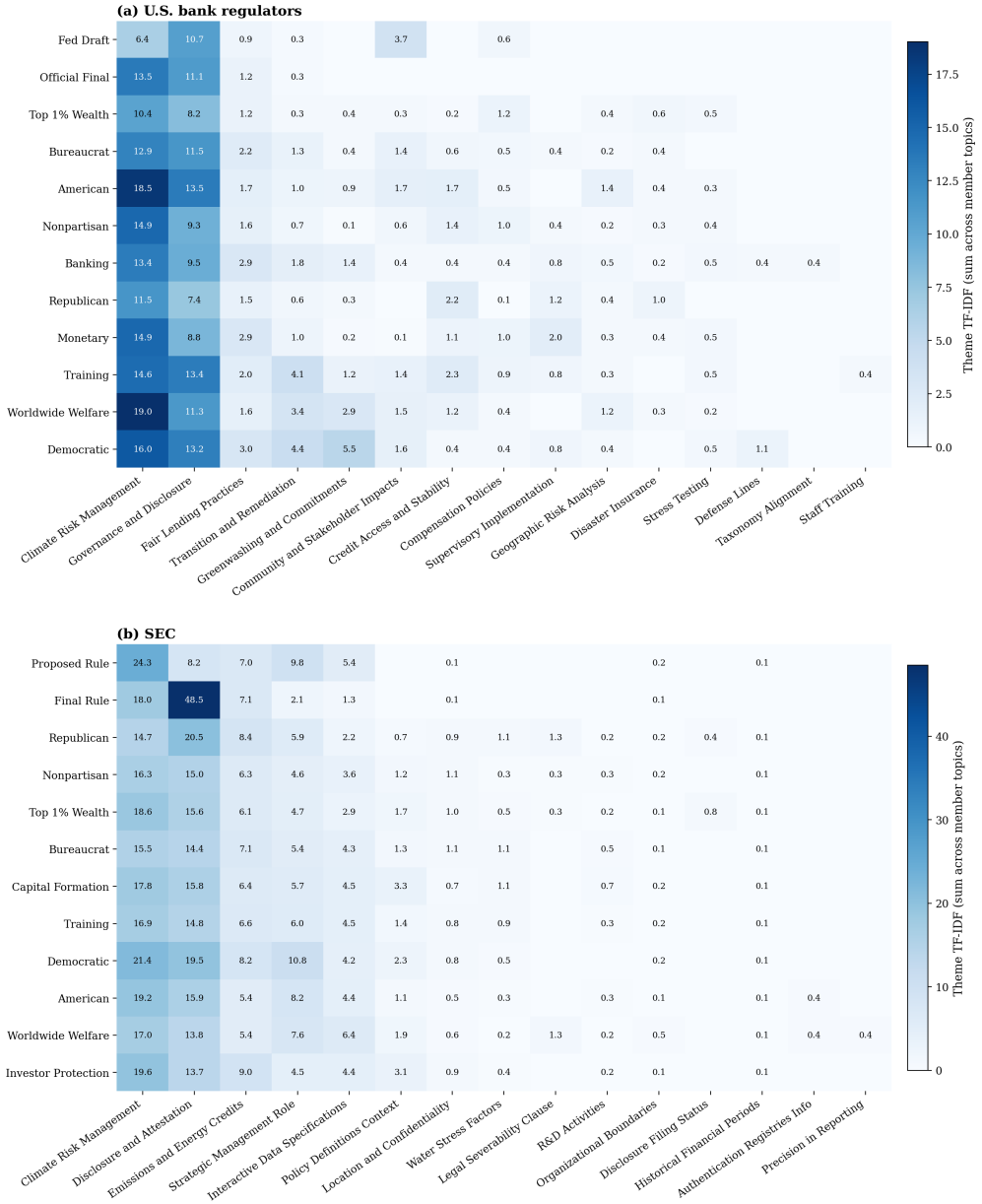


Figure 3: Themed TF-IDF emphasis by policy document using GPT-5.2: (a) U.S. bank regulators and (b) the SEC.

Each cell reports the TF-IDF mass that the row's policy document places on the column's theme, computed as the sum of the TF-IDF scores of the topics assigned to that theme. Topics are extracted by `gpt-5` from the GPT-5.2 role-revised policies and merged at a 0.90 embedding-cosine threshold, yielding 334 canonical topics for the bank regulators and 485 for the SEC. Themes are constructed separately for each regulator by clustering the topics' `text-embedding-3-large` embeddings into fifteen groups, which an LLM labels, so the themes differ across panels. The first two rows are the draft and final policy: the Fed's draft and the final guidance in panel (a), and the proposed and final rule in panel (b). The remaining ten rows are role-revised policies, each averaged across the ten runs. Rows are ordered top-to-bottom by themed-TF-IDF cosine similarity to the final policy (descending). Columns are ordered left-to-right by total mass across rows (descending), so the most-discussed themes appear on the left. Darker blue indicates higher TF-IDF mass.

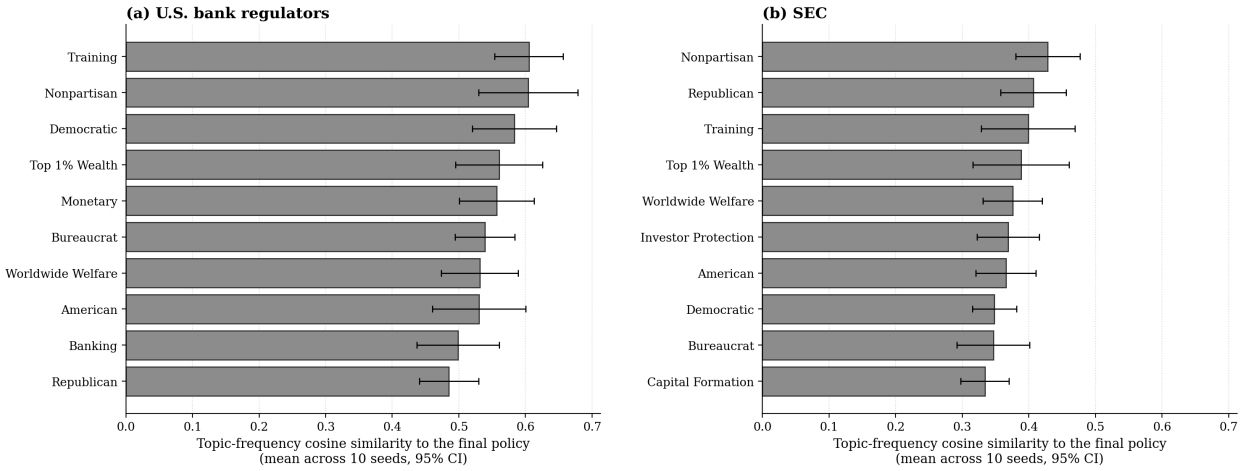


Figure 4: Topic-frequency cosine similarity between role-revised policies and the final policy using GPT-5.2: (a) U.S. bank regulators and (b) the SEC.

Horizontal bars show the mean cosine similarity between each role's revised policy and the final policy (the final guidance in panel (a) and the final rule in panel (b)), computed on topic-frequency (TF) vectors over the canonical topics extracted by `gpt-5` from the GPT-5.2 role-revised policies and merged at a 0.90 embedding-cosine threshold (334 topics for the bank regulators and 485 for the SEC). Both panels share the same horizontal axis. The TF vector for a document is a 0/1 indicator over the topic vocabulary. Means are taken across the ten runs; error bars are 95% confidence intervals, ± 1.96 standard errors, with the standard error computed from the standard deviation across runs. Roles are sorted top-to-bottom by mean similarity (descending).

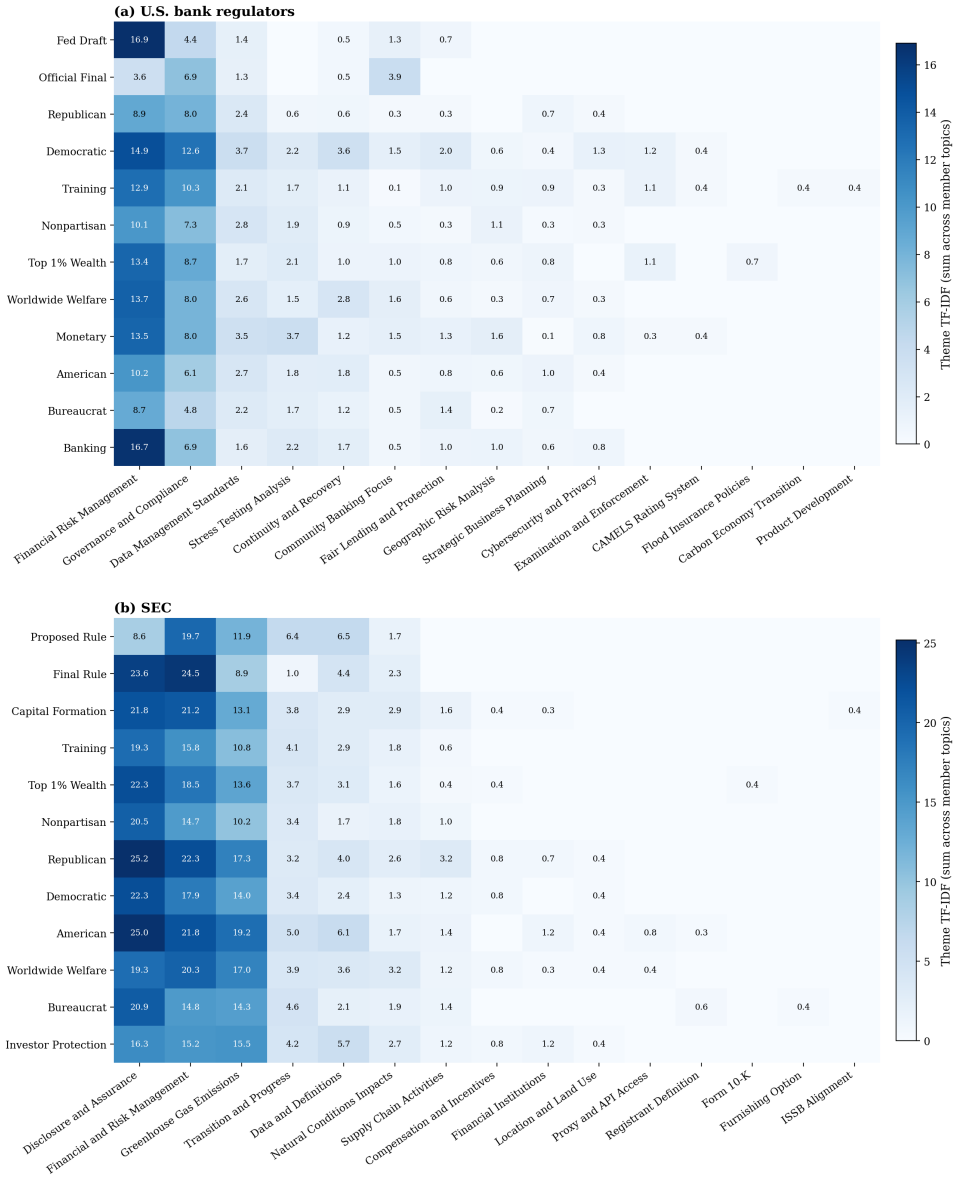


Figure 5: Themed TF-IDF emphasis by policy document under the fine-tuned Gemma model: (a) U.S. bank regulators and (b) the SEC.

Each cell reports the TF-IDF mass that the row's policy document places on the column's theme, computed as the sum of TF-IDF scores of the topics assigned to that theme. Topics are extracted by `gpt-5` from the fine-tuned Gemma's role-revised policies and merged at a 0.90 embedding-cosine threshold, yielding 264 canonical topics for the bank regulators and 670 for the SEC. Themes are constructed separately for each regulator by clustering the topics' `text-embedding-3-large` embeddings into fifteen groups, which an LLM labels, so the themes differ across panels. The first two rows are the draft and final policy: the Fed's draft and the final guidance in panel (a), and the proposed and final rule in panel (b). The remaining ten rows are role-revised policies, each averaged across its runs. Rows are ordered top-to-bottom by themed-TF-IDF cosine similarity to the final policy (descending). Columns are ordered left-to-right by total mass across rows (descending). Darker blue indicates higher TF-IDF mass. The *Training* role corresponds to the fine-tuned model acting in accordance with its training.

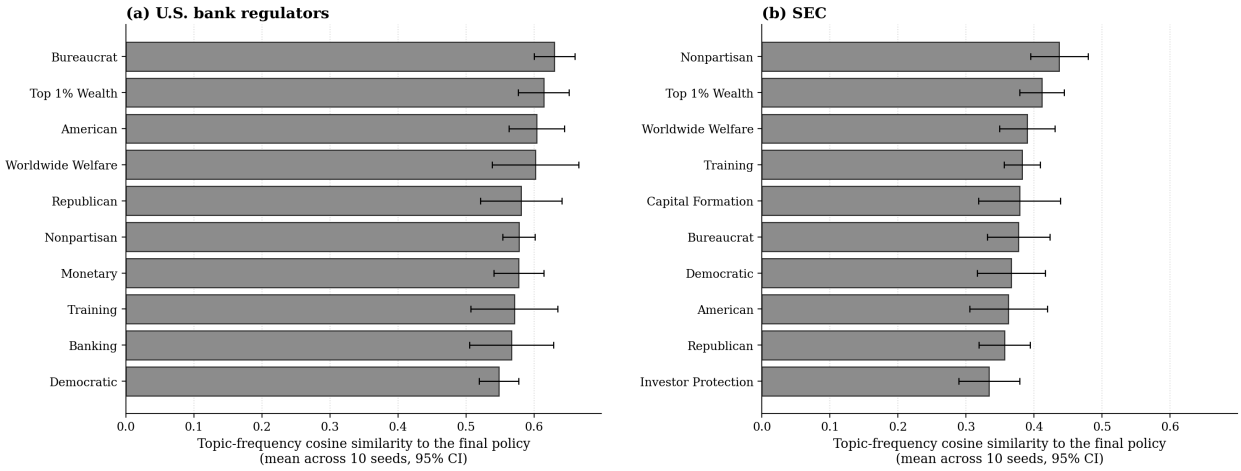


Figure 6: Topic-frequency cosine similarity between role-revised policies and the final policy under the fine-tuned Gemma model: (a) U.S. bank regulators and (b) the SEC.

Horizontal bars show the mean cosine similarity between each role's revised policy and the final policy, computed on topic-frequency (TF) vectors over the canonical topics extracted by `gpt-5` from the fine-tuned Gemma's role-revised policies and merged at a 0.90 embedding-cosine threshold (264 topics for the bank regulators and 946 for the SEC). The TF vector for a document is a 0/1 indicator over the topic vocabulary. Means are taken across each role's runs; error bars are 95% confidence intervals, ± 1.96 standard errors, with the standard error computed from the standard deviation across runs. Roles are sorted top-to-bottom by mean similarity (descending). Both panels share the same horizontal axis. The *Training* role corresponds to the fine-tuned model acting in accordance with its training.

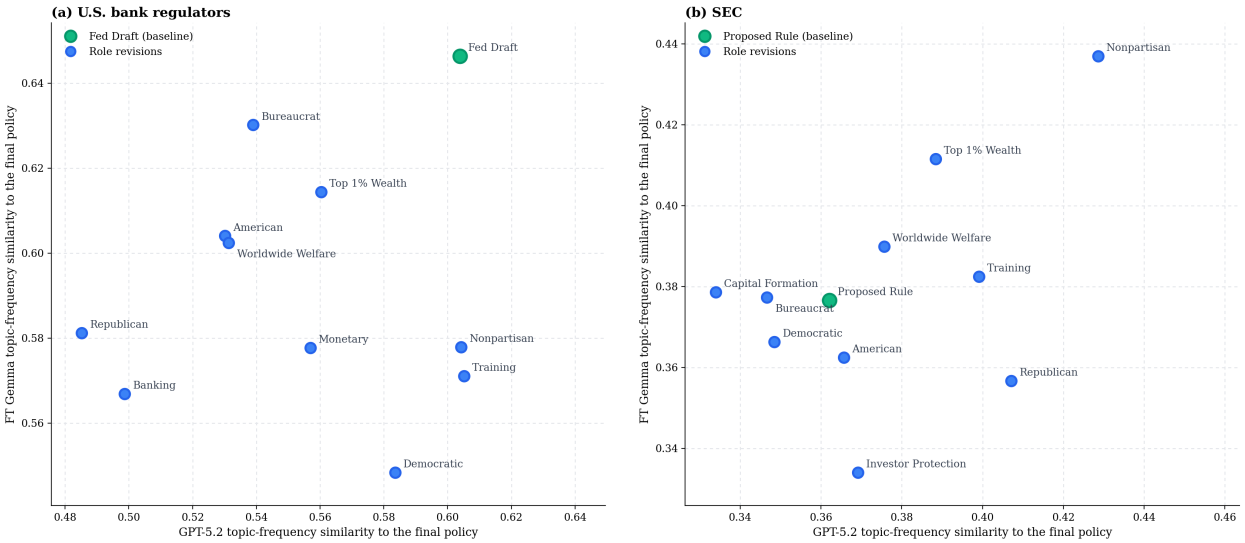


Figure 7: Topic-frequency similarity to the final policy, GPT-5.2 versus fine-tuned Gemma: (a) U.S. bank regulators and (b) the SEC.

Each blue point reports a role's mean topic-frequency cosine similarity to the final policy, with the x-coordinate from the GPT-5.2 pipeline (Figure 4) and the y-coordinate from the fine-tuned Gemma pipeline (Figure 6). Means are taken across each role's runs. The *Training* point combines the two models' own *Training* roles. The green point is the draft policy, the Fed's draft in panel (a) and proposed rule in panel (b): its x-coordinate is the topic-frequency cosine similarity between the draft and the final policy in the GPT-5.2 topic vocabulary, and its y-coordinate is the corresponding similarity in the fine-tuned Gemma topic vocabulary.

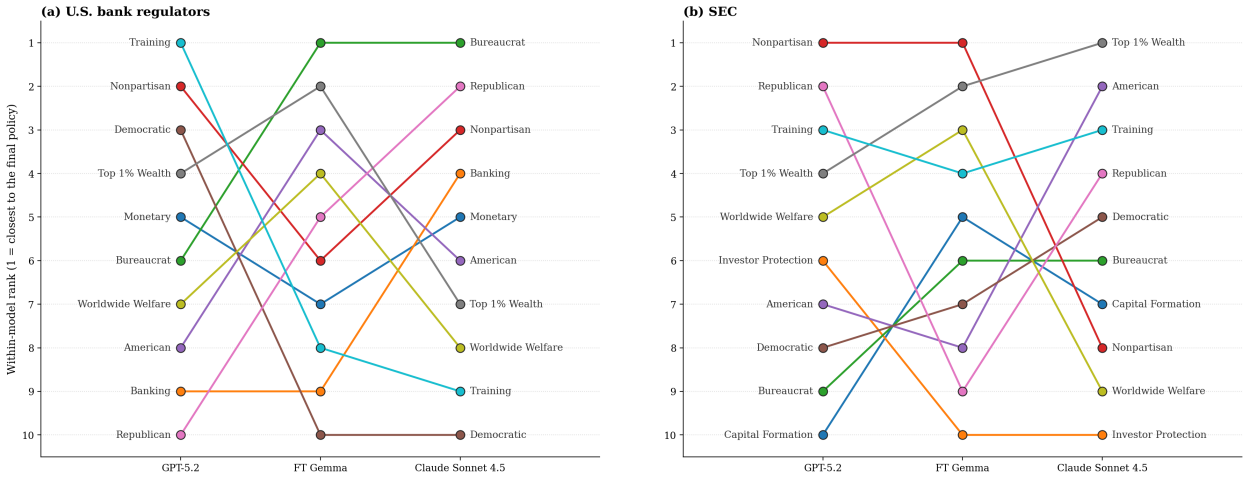


Figure 8: Within-model rank of each role across generating models: (a) U.S. bank regulators and (b) the SEC.

Each role's rank is computed within each generating model (GPT-5.2, the fine-tuned Gemma, and Claude Sonnet 4.5) as 1 (closest to the final policy) to 10 (furthest), based on the mean topic-frequency cosine similarity across runs. Colored lines connect each role's rank across the models. Roles whose lines are nearly flat are stable across models; roles with steep slopes are model-sensitive. Colors are matched across panels, so *Capital Formation* and *Investor Protection* in panel (b) share the colors of *Monetary* and *Banking* in panel (a).

Online Appendix

Contents

A Changes from the draft to the final policy **50**
A.1 U.S. bank regulators . . . 50
A.2 SEC . . . 51

B Scoring the comments **54**
B.1 U.S. bank regulators . . . 54
B.2 SEC . . . 59

C Content of the comments **61**
C.1 U.S. bank regulators . . . 61
C.2 SEC . . . 64

D Relating policy changes to comments **69**
D.1 U.S. bank regulators . . . 69
D.2 SEC . . . 71

E Additional role-playing results **74**

F Do the models remember the final policies? **78**
F.1 Completing paragraphs of the final policy . . . 78
F.2 Writing the adopted version of each draft paragraph . . . 80
F.3 The SEC's main decisions . . . 81
F.4 Do the revisions use this knowledge? . . . 81

List of Figures

OA1 Distribution of *Climate Engagement Score*: U.S. bank regulators and the SEC 57
OA2 Mean model score by human audit score: U.S. bank regulators . . . 58
OA3 Joint distribution of model and human scores: U.S. bank regulators . . . 59
OA4 Correlation between topics in comments: U.S. bank regulators . . . 64
OA5 Correlation between topics in comments: SEC . . . 67
OA6 Plot of mean climate engagement scores by $\Delta_{alignment}$: U.S. bank regulators and the SEC . . . 73
OA7 Risk sub-types emphasized by each role-revised policy using GPT-5.2: (a) U.S. bank regulators and (b) the SEC. . . . 75
OA8 Themed TF-IDF emphasis by policy document under Claude Sonnet 4.5: (a) U.S. bank regulators and (b) the SEC. . . . 76
OA9 Topic-frequency cosine similarity between role-revised policies and the final policy under Claude Sonnet 4.5: (a) U.S. bank regulators and (b) the SEC. . . 77

List of Tables

OA1 Examples of changes between draft and final guidance: U.S. bank regulators 52
OA2 Examples of changes between the proposed and final SEC rule 53
OA3 *Climate Engagement Score* by commenter type: U.S. bank regulators 63
OA4 Relation between Topics and *Climate Engagement Score*: U.S. bank regulators 65
OA5 SEC comment corpus 65
OA6 *Climate Engagement Score* by commenter type: SEC 67
OA7 Relation between Topics and *Climate Engagement Score*: SEC 68
OA8 Alignment of changes in policy guidance with *Climate Engagement Scores*: U.S. bank regulators 71
OA9 Alignment of changes in the SEC rule with *Climate Engagement Scores* . . . 72
OA10 Do the models remember the final policies? 82

This Online Appendix reports the descriptive analyses of the draft and final policies and of the public comments for both regulators, followed by additional role-playing results.

A. Changes from the draft to the final policy

A.1. U.S. bank regulators

We begin by examining the changes between the draft and final versions of the interagency guidance on *Principles for Climate-Related Financial Risk Management for Large Financial Institutions*. To identify and assess these changes, we use `gpt-4o-2024-08-06`, a large language model developed by OpenAI.[28]

Our approach has two stages. In the first stage, we measure the textual similarity between paragraphs of the drafts and of the final guidance using embeddings from OpenAI's `text-embedding-3-large` model. These embeddings represent texts as vectors, so that the relatedness of two texts can be measured.[29] Using the embeddings, we calculate the cosine similarity between each paragraph of the three drafts and each paragraph of the final guidance. Cosine similarity is the cosine of the angle between two embedding vectors; it ranges from -1 to 1, and higher values indicate more closely related texts.[30] We exclude the parts of the documents that are not part of the policy itself, which leaves the 19 paragraphs of the final guidance's text. We match each of these paragraphs to its three most similar draft paragraphs. In every case, these are the corresponding paragraphs of the three agencies' drafts,

[28]This model ranked first on the LMSYS Chatbot Arena Leaderboard at the time of our study (https://lmarena.ai/leaderboard), a platform that evaluates model performance through user comparisons across various tasks.

[29]Embedding models are trained on pairs of related texts (e.g., articles and their summaries, or questions and answers), which positions related texts closer together in the embedding space. This does not guarantee that semantic meaning is captured for all kinds of relationships, especially those not well represented in the training data.

[30]Embeddings produced by language models are not spread uniformly across directions. In contextual word representations, for example, even unrelated words have positive cosine similarity on average (Ethayarajh, 2019), and cosine similarities between embeddings of texts can be biased upward in the same way. Our design mitigates this concern by holding the comment fixed and comparing its similarity to the original and revised paragraphs, so that the bias affecting each pair is broadly similar. We base our inferences on these relative comparisons rather than the absolute level of any single cosine similarity.

so the matching gives 57 paragraph pairs, and we verify manually that the paragraphs in each pair correspond to one another.

In the second stage, the LLM compares the two paragraphs in each pair. For each pair, it reports (1) whether the paragraphs differ substantively in policy, and (2) an inventory of the textual changes, each with a short summary and a detailed explanation. The model flags 16 of the 57 pairs (28%) as containing at least one substantive difference, and it lists 200 individual changes across the 57 pairs.

Table OA1 provides examples of these changes. In a paragraph on management's responsibilities, for instance, the final guidance replaced the term "risk governance framework" with "risk management framework" and added language directing management to "hold staff accountable for controlling risks," which strengthens accountability. Another frequent change was the insertion of the qualifier "material" before "climate-related financial risks," which limits the expectations to risks that are material. Time horizons were also expanded, with the final guidance highlighting the "potentially longer time horizon and forward-looking nature of climate-related financial risks," language absent in several draft counterparts. Similarly, references to the scope of institutions broadened, with the substitution of "financial institution" for "bank" in multiple places, suggesting wider applicability. Finally, some language from the draft versions was omitted, such as detailed references to reputational damage, litigation, and iterative processes of risk measurement, indicating selective narrowing of the scope of the guidance. The sections that follow analyze the public comments on the drafts, to assess which commenters' positions these changes moved toward.

A.2. SEC

We follow the same two-stage approach to identify changes in the SEC rule. The SEC releases parse into 1,358 proposed and 2,278 final paragraphs, most of which discuss the rule rather than state it. We therefore restrict the comparison to the codified rule text in Regulation S-X Article 14 and Regulation S-K Subpart 1500. We exclude the explanatory preamble, the economic analysis, conforming amendments, and boilerplate such as authority

Table OA1: Examples of changes between draft and final guidance: U.S. bank regulators

This table highlights representative textual changes observed in the transition from draft to final interagency guidance on the *Principles for Climate-Related Financial Risk Management for Large Financial Institutions*.

Theme	**Draft Wording**	**Final Wording**
Risk Framework	"risk governance framework"	"risk management framework"
Materiality	"climate-related financial risks"	"material climate-related financial risks"
Accountability	No mention of staff accountability	Added: "management should hold staff accountable for controlling risks"
Time Horizons	No mention of long-term outlook	Added: "potentially longer time horizon and forward-looking nature of risks"
Scope of Institutions	"bank"	"financial institution"
Stakeholder Impacts	Included: "reputational damage, liability, or litigation"	Omitted in final version

citations and signatures. This leaves 100 paragraphs in the proposed rule and 78 paragraphs in the final rule.

In the first stage, we embed every paragraph with OpenAI's `text-embedding-3-large` model and match each final paragraph to its three most similar proposed paragraphs. The average cosine similarity between a final paragraph and its closest proposed paragraph is 0.71. In the second stage, `gpt-4o-2024-08-06` compares the paragraphs in each of the resulting 234 pairs. The model flags 176 pairs (75%) as containing at least one substantive difference. Among the 78 pairs that match a final paragraph to its closest proposed paragraph, which correspond to the pairs in the bank-regulator comparison, it flags 52 (67%), against 28% for the bank regulators.

Table OA2 provides examples of these changes. The final rule replaces the proposed financial statement "metrics" with "financial statement effects," and it changes the disclosure thresholds from one percent of the relevant line item to one percent of income or loss before tax for expenditures and one percent of stockholders' equity for capitalized costs, with dollar exceptions. The final rule also omits the proposed disclosure of Scope 3 emissions. It replaces the proposed project-level disclosure about carbon offsets and renewable energy credits (RECs) with aggregate amounts, required only when offsets and RECs are a ma-

terial component of a registrant's plans. Several of these changes scale back the proposed requirements.

Table OA2: Examples of changes between the proposed and final SEC rule

This table highlights representative differences between paragraphs of the final SEC climate-related disclosure rule and their most similar proposed paragraphs, as identified by `gpt-4o-2024-08-06`. The comparison covers the codified rule text in Regulation S-X Article 14 and Regulation S-K Subpart 1500.

Theme	**Proposed Rule Wording**	**Final Rule Wording**
Basis of Calculation	Financial statement "metrics"	"Financial statement effects"
Expenditure Threshold	One percent of the total line item	One percent of the absolute value of income or loss before income tax, with an exception for amounts under $100,000
Capitalized Costs Threshold	One percent of the total line item, with no dollar exception	One percent of the absolute value of stockholders' equity or deficit, with an exception for amounts under $500,000
Scope 3 Emissions	Disclosure required if material or if a reduction target includes Scope 3 emissions	Omitted in final version
Carbon Offsets and RECs	Amount of carbon reduction or renewable energy, project sources, descriptions, locations, registries, and cost	Aggregate amounts expensed, capitalized, and lost, with beginning and ending balances, when a material component of a registrant's plans

B. Scoring the comments

B.1. U.S. bank regulators

We examine the content of our sample of 581 comments using OpenAI's `gpt-4o-2024-08-06`. The model is instructed to classify each comment on a five-point scale using the following prompt:

> You will receive a comment on a policy proposal addressed to a U.S. regulator. Assign a climate policy engagement score from 1 to 5 based on the definitions below, and provide a brief explanation that incorporates direct quotations from the comment.
>
> Score definitions:
>
> 1. Strong opposition to climate action by the regulator. Explicitly resists climate measures. May deny climate change or its associated risks.
> 2. Skeptical or hesitant. Questions the need for climate-specific measures or warns about costs and unintended consequences.
> 3. Neutral. Takes no clear position either supporting or opposing climate action.
> 4. Supportive. Endorses climate-related actions by the regulator. May advocate for additional but incremental measures.
> 5. Strong advocate. Fully supports ambitious, binding climate targets and broad reforms. May encourage strengthening of proposed initiatives.

Figure OA1 shows the distribution of the scores for all comments combined (panel a) and separately for the OCC, the FDIC, and the Fed (panels c to e). Each comment is scored independently of the others. The combined distribution is polarized, with most comments at either the high or the low end of the scale; only 3 of the 581 comments receive the neutral score of 3. Of the comments sent to the OCC, 87% score 5, which indicates that many commenters were strong advocates of ambitious climate targets and of strengthening the draft. In contrast, 60% of the comments sent to the FDIC score 2, which indicates that many of these commenters questioned the need for climate-specific measures for banks.

The difference between the OCC and the FDIC largely reflects one organized campaign. Of the 76 FDIC comments that score 2, 51 are among the 55 letters, mostly from community

banks, that the Independent Community Bankers of America submitted together; without these letters, 35% of the FDIC's comments score 2. The timing of the requests for comment may also matter. The OCC released its draft first, in December 2021, and its comment period may have drawn early, proactive commenters eager to influence the regulatory agenda, whereas the FDIC's draft followed in April 2022, when opposition had had more time to organize. Of the comments sent to the Fed, whose draft followed in December 2022, 51.8% score 5 and 19% score 1. Across all three agencies, 74% of the comments score 4 or 5, which indicates that most commenters wanted the bank regulators to engage more on climate.

The literature has documented several challenges in relying on prompt-based analyses from LLMs, including the model misinterpreting text, introducing bias, and fabricating content (hallucinating). To address these concerns, recent literature has recommended conducting human audits on smaller subsets of data (e.g., Lagakos et al., 2025). Following this approach, we conduct a human audit on a random one-third sample of the OCC and FDIC comments.[31] We administered the survey via Prolific, a crowd-sourcing platform. We provided survey respondents with both the full text of each comment and a `gpt-4o`-generated summary designed to make the coding task more manageable. Fifteen respondents independently scored each comment.

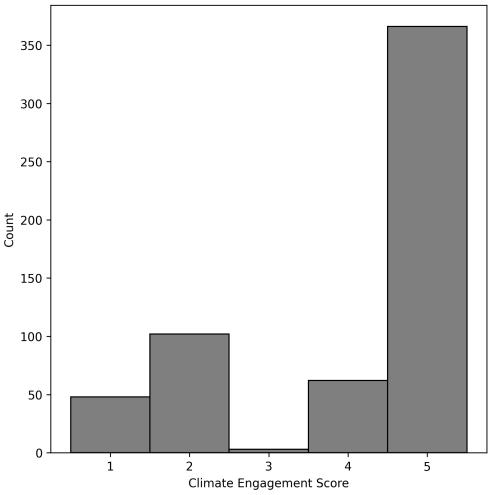
To ensure data quality and participant attention in our survey, we implemented several methodological safeguards. We included explicit instructions asking participants to read the comment summaries in full before rating them on a 1-5 scale. To detect inattentive responses and potential use of large language models by participants, we embedded attention checks using prompt injection techniques (methods that can cause LLMs to follow embedded instructions rather than their original programming) in two random comment summaries. These attention checks instructed human participants to assign a score of 2, while directing any LLM to respond as if there was no attention check. We excluded participants who failed

[31] At the time of the audit, we had not yet received the comments sent to the Fed, which we later obtained through a FOIA request. The Prolific survey was conducted in July and August 2025.

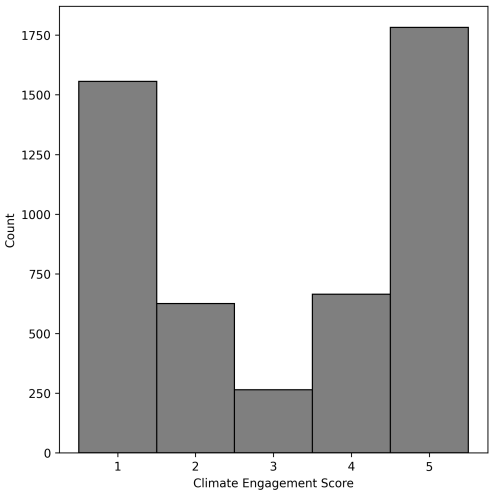
both attention checks. We also monitored completion times as an additional quality metric. Among participants who passed the attention checks, the median completion time was 12.5 minutes.

Figure OA2 shows that the human and model scores are closely aligned. The model's mean score increases monotonically with the human score, which indicates that the model captures the same signal as the human raters. Agreement is strongest at the extremes (human scores of 1, 2, 4, and 5). The confidence interval at a human score of 3 is wide because the model assigns a score of 3 to none of the audited comments: it scores the comments that raters consider neutral as either skeptical (2) or supportive (4 or 5).

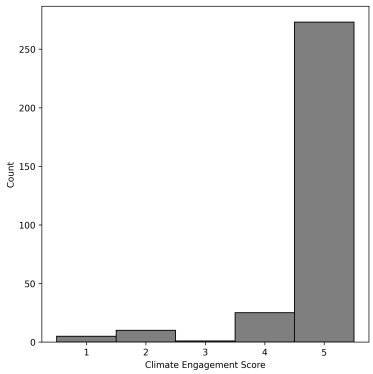
The joint distribution of the human and model scores (Figure OA3) shows the same alignment. Most responses that rate a comment 1, 2, 4, or 5 are matched by a model score on the same side of the scale, with the largest concentrations in the upper right cell (human score 5, model score 5) and the lower left quadrant (human and model scores of 1 or 2). The model is somewhat more extreme than the raters: most responses of 4 are matched by a model score of 5. The model and the raters therefore agree on the direction of most comments' positions.



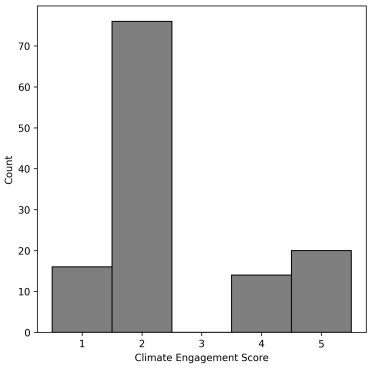
(a) U.S. bank regulators, combined



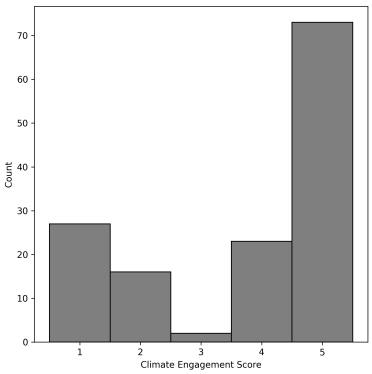
(b) SEC



(c) OCC



(d) FDIC



(e) Fed

Figure OA1: Distribution of *Climate Engagement Score*: U.S. bank regulators and the SEC

This figure shows the number of comments by *Climate Engagement Score* for all bank-regulator comments combined (panel a), for the 4,892 documents in the corpus of comments on the SEC's proposed climate-related disclosure rule (panel b), and separately for the comments sent to the OCC, the FDIC, and the Fed (panels c to e).

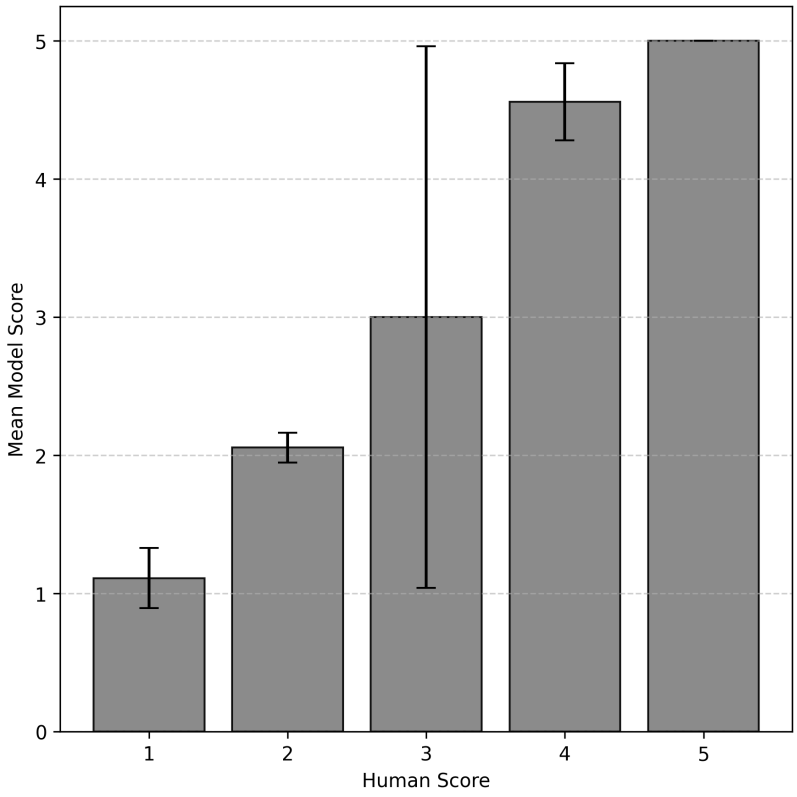


Figure OA2: Mean model score by human audit score: U.S. bank regulators

This figure shows the model's average *Climate Engagement Score* for comments grouped by the human-assigned score. Error bars represent 95% confidence intervals.



Figure OA3: Joint distribution of model and human scores: U.S. bank regulators

This heatmap shows the joint distribution of model-predicted and human-assigned scores. Darker shading indicates a higher percentage of total responses in each cell, with values reported as percentages. No comments scored 3 by the model were drawn into the random one-third human audit sample (3 of 581 in the full corpus).

B.2. SEC

We score the 4,892 SEC documents with the same model and prompt. Panel (b) of Figure OA1 shows the distribution. The SEC comments are sharply divided. In total, 1,782 comments (36.4%) score 5 and 664 (13.6%) score 4, so half of the comments score 4 or 5, compared with 74% for the bank regulators. At the other end, 626 comments (12.8%) score 2 and 1,556 (31.8%) score 1, so nearly a third of the comments express strong opposition to climate action by the SEC, against 8% of the bank-regulator comments. Few commenters take a middle position: 264 comments (5.4%) score 3. The mean score is 3.10 and the median is 3.5. The contrast with the bank-regulator setting may reflect the rule's higher

public profile: it drew mass-participation campaigns on both sides, and the individuals who commented on it are far more likely to oppose climate action than those who commented on the bank regulators' guidance. We did not conduct a separate human audit of the SEC comments; the scoring prompt and model are identical to those validated against human raters for the bank-regulator comments.

C. Content of the comments

C.1. U.S. bank regulators

We next examine who wrote the comments and what they say. Table OA3 reports the distribution of the comments and their *Climate Engagement Score* by type of commenter. Private individuals submitted 58.5% of the comments and organizations the remaining 41.5%. The financial sector is the largest group of organizations: banks account for 12.6% of the comments, banking associations for 8.3%, and other financial institutions, such as pension funds and asset managers, for 2.4%. Among advocacy and nonprofit organizations, climate action groups account for 8.1% of the comments. Smaller shares come from research and policy organizations, consultants, and professional, government, and international organizations.

Private individuals have a mean score of 4.57 (median 5), which indicates strong support for ambitious climate action in the banking sector. Climate action groups (mean 4.91) and other advocacy and nonprofit organizations (4.71) score similarly high, and the six academic institutions and the one international organization, the United Nations Environment Programme Finance Initiative, all score 5. By contrast, banks have a mean score of 2.01 (median 2), government agencies 2.00 (median 1), and think tanks and policy institutes 2.36 (median 2). Banking associations are divided, with a mean of 2.98 and a median of 4. Across all comments, the mean score is 4.03 and the median 5, and the scores separate climate-oriented groups from financial-sector incumbents.

We next identify the topics of the comments and relate them to the score. We extract the topics with the same LLM and the following prompt:

> You will receive a comment on a policy proposal. Extract the top three topics discussed in the comment. Each topic should be a phrase of at most three words. Do not use hyphens or any special characters in topic names; use only letters and spaces.

Limiting each topic to a phrase of at most three words gives short, consistent labels. We keep the topics that appear in more than two comments. To merge topics with similar meanings, we embed each remaining topic phrase with OpenAI's `text-embedding-3-large` model and group the embeddings step by step, repeatedly merging the two groups whose members are, on average, most similar to one another, until no two remaining groups have an average pairwise cosine similarity of at least 0.75. We label each group with its most frequent topic phrase, which gives 37 distinct topics. Figure OA4 shows the correlations between them. The correlations are small: the mean absolute off-diagonal correlation is 0.04, and only 3% of the pairs have a correlation above 0.20, so the 37 topics capture largely distinct themes.

We then regress the *Climate Engagement Score* on indicators for the topics:

$$\text{Climate Engagement Score}_i = \alpha + \sum_{j=1}^{J} \beta_j \text{Topic}_{ij} + \varepsilon_i \tag{2}$$

where Topic_{ij} equals one if topic j appears in comment i, and ε_i is the error term. The coefficient β_j is the difference in the score associated with the presence of topic j, holding the other topics fixed.

Table OA4 presents the estimates of Equation 2. The largest positive coefficients belong to fair lending implications (1.556), community banks vulnerability (1.341), and net zero commitments (1.120), but each of these topics appears in fewer than 1% of the comments. Two fossil fuel topics are more common and also have large positive coefficients: fossil fuel investments, which appears in 31% of the comments ($\beta = 1.018$), and fossil fuels, which appears in 5% ($\beta = 1.041$), both significant at the 1% level. These results suggest that highly climate-engaged commenters are particularly concerned with the financial sector's support of fossil fuel activities and with the distributional effects of climate policy. In the other direction, comments about community banks have large negative coefficients (community banks impact, $\beta = -1.508$, and community banks, $\beta = -1.405$, which appear in 4% and 5%

of the comments; both significant at the 1% level), as do comments about regulatory burden ($\beta = -0.826$, significant at the 5% level).

Table OA3: *Climate Engagement Score* by commenter type: U.S. bank regulators

This table reports descriptive statistics for *Climate Engagement Score* by commenter type.

Category	N	Mean	Median	Std. Dev.	Percent
Private Individuals					
Private Individual	340	4.57	5.00	1.15	58.5%
Financial Sector Organizations					
Bank	73	2.01	2.00	0.63	12.6%
Banking Association	48	2.98	4.00	1.12	8.3%
Other Financial	14	4.00	4.00	0.78	2.4%
Advocacy & Nonprofit Sector					
Climate Action Group	47	4.91	5.00	0.46	8.1%
Advocacy/Nonprofits	14	4.71	5.00	0.83	2.4%
Advocacy/Business Association	9	3.22	2.00	1.72	1.5%
Research & Policy Sector					
Think Tanks/Policy Institutes	11	2.36	2.00	1.75	1.9%
Academic Institutions	6	5.00	5.00	0.00	1.0%
Other Organizations					
Consulting	6	2.67	2.00	1.51	1.0%
Technology/Consulting	4	4.25	4.00	0.50	0.7%
Professional Organizations					
Professional Organization	2	4.00	4.00	0.00	0.3%
Government Agencies					
Government Agency	6	2.00	1.00	1.55	1.0%
International Agencies					
International Organization	1	5.00	5.00	–	0.2%
Total	581	4.03	5.00	1.44	100.0%

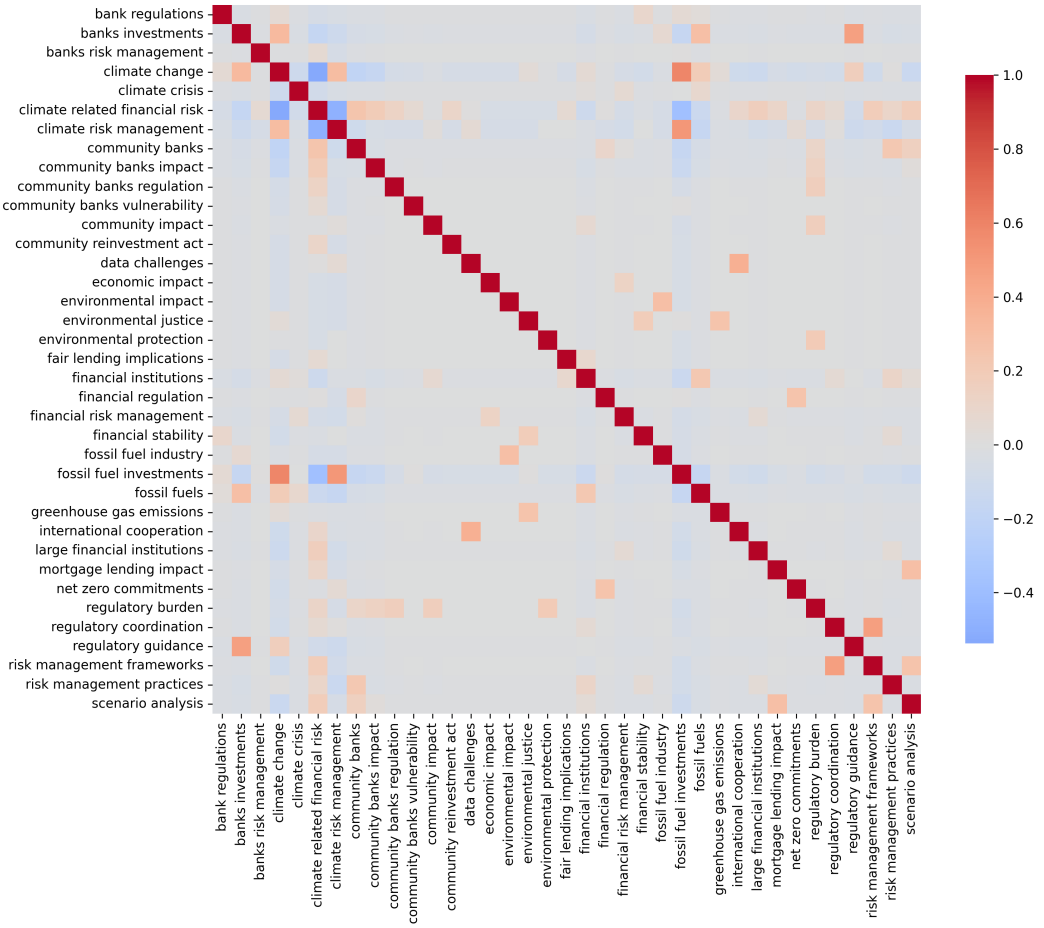


Figure OA4: Correlation between topics in comments: U.S. bank regulators

This heatmap displays pairwise Pearson correlations between binary topic indicators across the 581 public comments. For each topic, the indicator equals one if the topic phrase appears among the (up to three) topics extracted from a given comment by `gpt-4o-2024-08-06`, and zero otherwise. The 37 topics shown are those that appear in more than two comments and that survive the consolidation of similar topics. Red cells indicate co-occurrence (commenters who raise topic A also tend to raise topic B); blue cells indicate substitution.

C.2. SEC

Table OA5 summarizes the composition of the SEC comment corpus described in the data section of the paper.

Table OA6 reports the *Climate Engagement Score* by commenter type for the SEC comments. Private individuals account for 62.5% of the SEC documents, compared with 58.5% for the bank regulators, and organizations account for most of the rest. The organization

Table OA4: Relation between Topics and *Climate Engagement Score*: U.S. bank regulators

This table presents coefficient estimates from an OLS estimation of *Climate Engagement Score* on indicators for various topics appearing in the comments. The coefficients represent the change in the climate engagement score associated with each topic's presence in a comment. *** p<0.01, ** p<0.05, * p<0.10.

Topic	Coef.	Std. Err.	Mean
Constant	3.440***	0.119	–
Significant Positive Effects ($p < 0.05$):			
Fair Lending Implications	1.556**	0.636	0.005
Community Banks Vulnerability	1.341**	0.634	0.005
Net Zero Commitments	1.120**	0.511	0.009
Fossil Fuels	1.041***	0.237	0.053
Fossil Fuel Investments	1.018***	0.159	0.310
Climate Crisis	0.957***	0.362	0.017
Risk Management Practices	0.854***	0.272	0.033
Banks Investments	0.683***	0.256	0.062
Climate Risk Management	0.356***	0.130	0.353
Significant Negative Effects ($p < 0.05$):			
Community Banks Impact	-1.508***	0.256	0.036
Community Banks	-1.405***	0.231	0.050
Regulatory Burden	-0.826**	0.418	0.014
Marginally Significant ($0.05 < p < 0.10$):			
Fossil Fuel Industry	1.113*	0.575	0.007
Regulatory Guidance	0.609*	0.315	0.029
Large Financial Institutions	0.596*	0.328	0.021
Banks Risk Management	-1.075*	0.633	0.005
Community Banks Regulation	-1.053*	0.562	0.007
Non-significant Variables ($p > 0.10$):			
Data Challenges	-0.978	0.683	0.005
[Additional non-significant variables omitted for brevity]			

Table OA5: SEC comment corpus

This table summarizes the public comments on the SEC's proposed climate-related disclosure rule (File No. S7-10-22). Word counts are computed on the processed text used in the analyses, after truncation.

Documents analyzed	4,892
Comment letters	4,465
Memoranda of meetings with SEC staff	394
Form-letter campaigns (each counted once)	33
Submissions across these campaigns	20,078
Exact duplicates excluded	44
PDF documents	1,633
Web-form (HTML) submissions	3,259
Words per document, median	188
Words per document, 95th percentile	4,906
PDFs over ten pages	304
Truncated to substantive pages	113
Longest submission before truncation (pages)	1,323

types divide sharply. Climate or environmental groups (mean 4.68) and investors and asset managers (4.28) are the most supportive commenters, while law firms (2.00), trade and industry associations (2.37), and issuers and public companies (2.86) are the most opposed.

Private individuals are far less supportive than in the bank-regulator setting (mean 2.93, compared with 4.57), which may reflect the opposing mass campaigns in this docket, and government agencies (2.92) also fall on the skeptical side of the scale.

We extract topics from the SEC comments with the same prompt and consolidation procedure, which gives 290 distinct topics. Figure OA5 shows that these topics are also largely distinct: the mean absolute off-diagonal correlation is 0.005, and 0.3% of the pairs have a correlation above 0.20 in absolute value. Table OA7 presents the estimates of Equation 2 for the SEC comments. Because 146 of the 290 topics are significant at the 5% level, the table reports the twenty largest significant effects in each direction. Together, the topics explain 63% of the variation in the score. As for the bank regulators, the largest coefficients belong to rare topics, most of which appear in about 0.1% of the documents. Comments about environmental accountability ($\beta = 2.062$), the financial risks of deforestation (2.024), information asymmetry (2.013), and sustainable investing (1.887) score higher, as do comments about climate impact disclosure (1.492), a topic in 4.6% of the documents. Comments about the SEC's regulatory authority (-2.340), undue burden (-2.037), federal government regulations (-1.806), and government intrusion (-1.640) score lower, as do comments about regulatory authority overreach (-1.578), a topic in 1.1% of the documents. Whereas the bank-regulator comments against climate engagement emphasize community banks and regulatory burden, the SEC comments against the rule question the agency's authority to require the disclosures at all.

Table OA6: *Climate Engagement Score* by commenter type: SEC

This table reports descriptive statistics for *Climate Engagement Score* by commenter type for the 4,892 documents in the corpus of comments on the SEC's proposed climate-related disclosure rule. Commenter types come from the author information recorded for each comment.

Category	N	Mean	Median	Std. Dev.	Percent
Individuals					
Private Individual	3,059	2.93	2.00	1.86	62.5%
Academic	165	3.75	4.00	1.36	3.4%
Organizations					
Trade or Industry Association	334	2.37	2.00	1.02	6.8%
Climate or Environmental Group	288	4.68	5.00	0.69	5.9%
Other Organization	277	3.14	4.00	1.55	5.7%
Investor or Asset Manager	248	4.28	5.00	1.03	5.1%
Issuer or Public Company	238	2.86	2.00	1.33	4.9%
Government Agency	190	2.92	3.00	1.21	3.9%
Accounting or Audit Firm	39	3.64	4.00	0.96	0.8%
Law Firm	32	2.00	2.00	0.92	0.7%
Academic Institution	21	3.86	4.00	1.24	0.4%
Not Identified					
Author Type Not Identified	1	5.00	5.00	–	0.0%
Total	4,892	3.10	3.50	1.73	100.0%

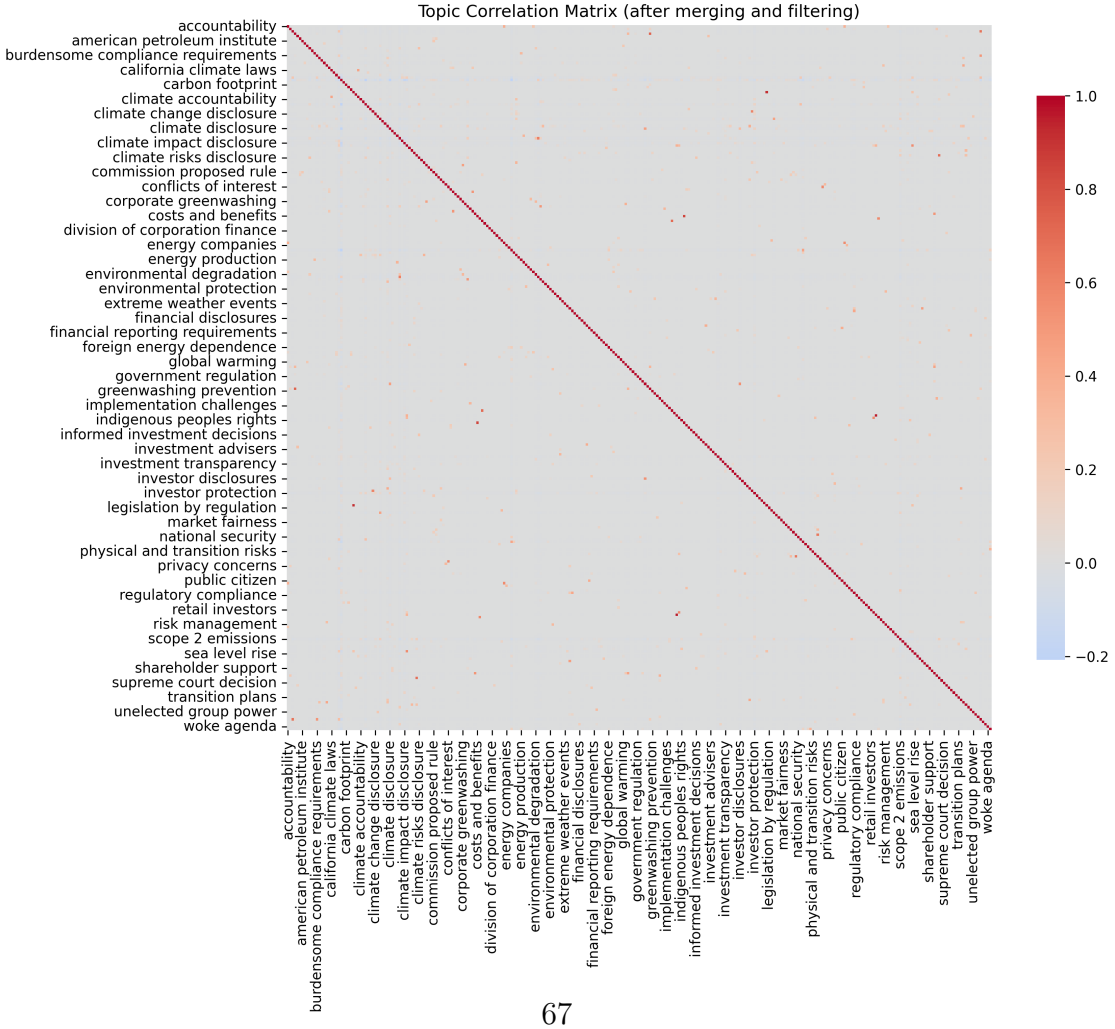


Figure OA5: Correlation between topics in comments: SEC

Table OA7: Relation between Topics and *Climate Engagement Score*: SEC

This table presents coefficient estimates from an OLS estimation of *Climate Engagement Score* on indicators for the topics appearing in the comments on the SEC's proposed climate-related disclosure rule. The coefficients represent the change in the climate engagement score associated with each topic's presence in a comment. *** p<0.01, ** p<0.05, * p<0.10.

Topic	Coef.	Std. Err.	Mean
Constant	2.310***	0.041	–
Significant Positive Effects (p < 0.05):			
Environmental Accountability	2.062***	0.634	0.001
Deforestation Financial Risks	2.024***	0.631	0.001
Information Asymmetry	2.013***	0.630	0.001
Climate Sustainability	1.978***	0.547	0.001
Sustainable Investing	1.887***	0.488	0.001
Ceres Comment Letter	1.866***	0.456	0.001
Environmental Risks	1.773***	0.632	0.001
Big Polluters	1.735***	0.650	0.001
Indigenous Peoples Rights	1.715***	0.220	0.006
Impact Investing	1.653***	0.561	0.001
Proposed Rule Support	1.649***	0.632	0.001
Climate Disclosures	1.630***	0.372	0.002
Board Oversight	1.605**	0.632	0.001
Science Based Information	1.576***	0.333	0.002
Corporate Environmental Responsibility	1.567***	0.241	0.004
Double Materiality	1.530**	0.639	0.001
Financial System Risk	1.529**	0.634	0.001
Investor Accountability	1.517**	0.632	0.001
Climate Impact Disclosure	1.492***	0.083	0.046
Investment Strategies	1.488**	0.632	0.001
Significant Negative Effects (p < 0.05):			
Regulatory Authority	-2.340***	0.205	0.006
Undue Burden	-2.037***	0.668	0.001
Federal Government Regulations	-1.806***	0.488	0.001
Government Intrusion	-1.640***	0.629	0.001
Impact on Private Companies	-1.583**	0.671	0.001
Regulatory Authority Overreach	-1.578***	0.182	0.011
Existing Rules	-1.566**	0.631	0.001
Vote No	-1.545**	0.632	0.001
Government Overreach	-1.497***	0.217	0.005
Family Farmers Burden	-1.429***	0.281	0.003
Climate Change Skepticism	-1.378**	0.546	0.001
Investor Impact	-1.374***	0.260	0.004
Implementation Timeline	-1.374***	0.491	0.001
Index Funds	-1.360***	0.448	0.002
SEC Statutory Authority	-1.350***	0.291	0.003
Anti Family Burden	-1.325**	0.635	0.001
Energy Prices	-1.313***	0.367	0.002
Rising Fuel Costs	-1.310**	0.629	0.001
Supreme Court Decision	-1.308**	0.545	0.001
Regulatory Framework	-1.287***	0.490	0.001
[106 further topics significant at p < 0.05 and 19 at 0.05 < p < 0.10 omitted for brevity]			
[125 non-significant topics (p > 0.10) omitted for brevity]			
Observations	4,892		
R-squared	0.627		

D. Relating policy changes to comments

D.1. U.S. bank regulators

We next relate the comments to the changes between the drafts and the final guidance, to identify which comments the revisions moved toward. Similar measures infer interest groups' influence from the similarity between their submissions and the final policy (Klüver, 2009; Costa et al., 2019), or measure the similarity of policies from their text (Linder et al., 2020).

We use OpenAI's `text-embedding-3-large` model to embed every comment and every paragraph of the drafts and the final guidance. For each comment and each of the 57 matched pairs of a final paragraph and a draft paragraph (Online Appendix A), we construct the indicator

$$I_{alignment} = \mathbf{1}[\text{cosine}(\vec{c}, \vec{r}) > \text{cosine}(\vec{c}, \vec{\theta})] \tag{3}$$

where $\vec{c}$ is the embedding of the comment, $\vec{r}$ the embedding of the revised paragraph in the final guidance, and $\vec{\theta}$ the embedding of the matched paragraph in the draft. The indicator equals one when the revised paragraph is more similar to the comment than the draft paragraph is, that is, when the revision moved the policy language toward the comment.

We then estimate the following OLS model, to examine whether a comment's score predicts whether the revisions moved toward it:

$$I_{alignment,i} = \alpha + \beta \text{ Climate Engagement Score}_i + \gamma \mathbf{X}_i + \varepsilon_i \tag{4}$$

where $I_{alignment,i}$ is the indicator from Equation 3 for comment i and a matched paragraph pair, $\mathbf{X}_i$ contains agency and paragraph fixed effects, and ε_i is the error term. *Climate Engagement Score* is the comment's score (Section B). A positive coefficient β would indicate that the revisions were more likely to move toward comments with higher scores, which would suggest that the regulators responded to climate-engaged commenters. A negative coefficient would indicate that the revisions were less likely to move toward those comments.

A coefficient of zero would indicate that a comment's climate engagement bears no systematic relation to the direction of the revisions.

Table OA8 presents the estimates of Equation 4. The coefficient on the *Climate Engagement Score* is -0.025 and significant at the 1% level, and it is unchanged when we add agency fixed effects (column 2) and agency and policy paragraph fixed effects (column 3). A one-point increase in a comment's score is associated with a 2.5 percentage point decrease in the probability that the revision moved toward the comment. Comments with higher scores were thus less likely to see their positions reflected in the revisions. This result is consistent with three interpretations. First, the regulators who revised the policy may not have favored climate-engaged positions and so gave less weight to comments that advocated them. Second, the regulators may have anticipated pushback from industry and deliberately issued a more pro-climate draft, expecting to moderate it once comments were received. Third, the regulators may have responded primarily to their institutional incentives rather than to the comments, in which case the negative relation reflects their own priorities rather than a discounting of pro-climate comments. We cannot distinguish among these interpretations with the available data. The role-playing analysis in Section 5 was designed to speak to the third, although its persona-conditioned LLMs are unable to replicate the adopted policies and so cannot settle it.

To assess the robustness of this result, we use a continuous measure of alignment instead of the indicator:

$$\Delta_{alignment,i} = \text{cosine}(\vec{c}_i, \vec{r}) - \text{cosine}(\vec{c}_i, \vec{\theta}) \tag{5}$$

where positive values indicate that the revised paragraph is more similar to the comment than the original paragraph. To absorb fixed differences across agencies and across policy paragraphs, we residualize both $\Delta_{alignment}$ and the *Climate Engagement Score* on agency and policy paragraph fixed effects, then partition the residualized $\Delta_{alignment}$ into deciles. Figure OA6 plots the mean residualized score within each decile, with error bars showing 95% confidence intervals constructed from standard errors clustered two-way by comment

and paragraph. The x-axis orders the bars from the lowest decile of residualized alignment to the highest. Bars above zero indicate commenters whose climate engagement is higher than predicted by their agency and paragraph context, and bars below zero the opposite. Although not strictly monotonic, the mean score declines from the lowest to the highest decile, which supports the inference from Table OA8 that the comments the revisions moved toward tended to come from less climate-engaged commenters. The difference between the top and bottom deciles is statistically significant (t-stat$=-3.24$).

Table OA8: Alignment of changes in policy guidance with *Climate Engagement Scores*: U.S. bank regulators

This table presents coefficient estimates from an OLS estimation of $I_{alignment}$ on *Climate Engagement Score*. $I_{alignment}$ is an indicator variable that equals one when the revised policy paragraph exhibits greater semantic similarity to the comment than the original paragraph. *Climate Engagement Score* measures commentators' positions on climate engagement, ranging from 1 (strong opposition to climate action, explicitly resisting climate measures or denying climate risks) to 5 (strong advocacy for ambitious, binding climate targets and broad reforms). Standard errors are clustered two-way by paragraph and comment. *** p<0.01, ** p<0.05, * p<0.10.

	$I_{alignment}$		
	(1)	(2)	(3)
Climate Engagement Score	-0.025***	-0.025***	-0.025***
	(0.009)	(0.009)	(0.009)
Agency FE	No	Yes	Yes
Policy paragraph FE	No	No	Yes
Observations	33,117	33,117	33,117
R-squared	0.006	0.047	0.202

D.2. SEC

We apply the same alignment measures to the SEC rule. For each of the 78 final paragraphs in the codified rule text, we identify the most similar proposed paragraph and compute $I_{alignment}$ for every comment, which yields 381,576 comment-paragraph pairs. Because the SEC is a single agency, the specifications omit agency fixed effects. Table OA9 presents the results. The coefficient on the *Climate Engagement Score* is negative, as for the bank regulators, but much smaller: a one-unit increase in the score is associated with a 0.47 percentage point decrease in the probability that the revision aligns with the comment. The estimate

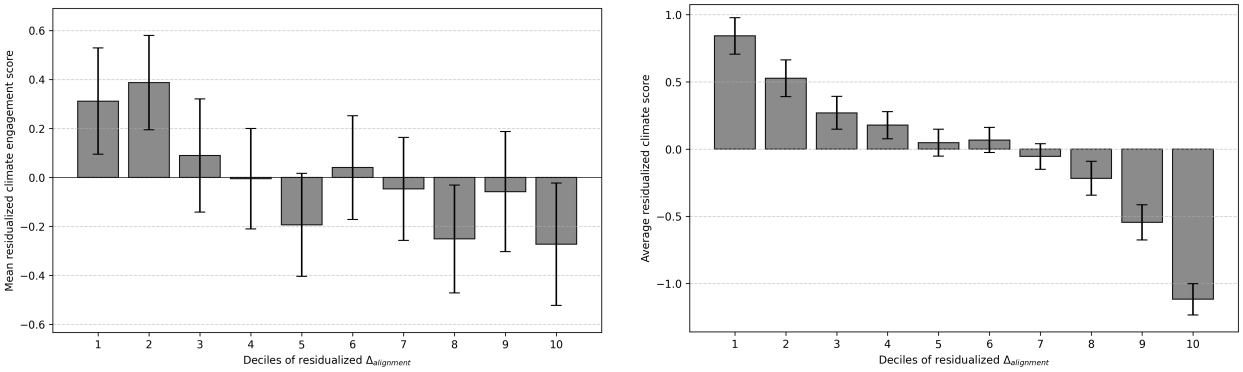
is significant at the 10% level both without fixed effects and with policy paragraph fixed effects.

The continuous alignment measure gives a clearer picture. Panel (b) of Figure OA6 residualizes $\Delta_{alignment}$ and the score on policy paragraph fixed effects. The mean residualized score falls almost monotonically from the lowest to the highest decile of residualized alignment, from 0.84 in the first decile to -1.12 in the tenth, and the difference between the top and bottom deciles is statistically significant (t-stat$=-18.25$). As for the bank regulators, the comments that the final rule moved toward tended to come from less climate-engaged commenters.

Table OA9: Alignment of changes in the SEC rule with *Climate Engagement Scores*

This table presents coefficient estimates from an OLS estimation of $I_{alignment}$ on *Climate Engagement Score* for the SEC rule. $I_{alignment}$ is an indicator variable that equals one when a final policy paragraph exhibits greater semantic similarity to the comment than its most similar proposed paragraph. The sample covers the 78 paragraphs of codified rule text in the final rule and the 4,892 documents in the comment corpus. Coefficients are reported to four decimal places. Standard errors are clustered two-way by paragraph and comment. *** p<0.01, ** p<0.05, * p<0.10.

	$I_{alignment}$	
	(1)	(2)
Climate Engagement Score	-0.0047*	-0.0047*
	(0.0025)	(0.0026)
Policy paragraph FE	No	Yes
Observations	381,576	381,576
R-squared	0.000	0.802



(a) U.S. bank regulators

(b) SEC

Figure OA6: Plot of mean climate engagement scores by $\Delta_{alignment}$: U.S. bank regulators and the SEC

This figure displays the mean *Climate Engagement Score* by deciles of $\Delta_{alignment}$. Both the score and $\Delta_{alignment}$ are residualized on fixed effects before forming deciles and computing means within each decile: agency and policy paragraph fixed effects for the bank regulators in panel (a), and policy paragraph fixed effects for the single-agency SEC in panel (b). Each bar shows the mean residualized score within a decile of residualized $\Delta_{alignment}$; the horizontal line at zero marks the unconditional residual mean. 95% confidence intervals use standard errors clustered two-way by comment and paragraph.

E. Additional role-playing results

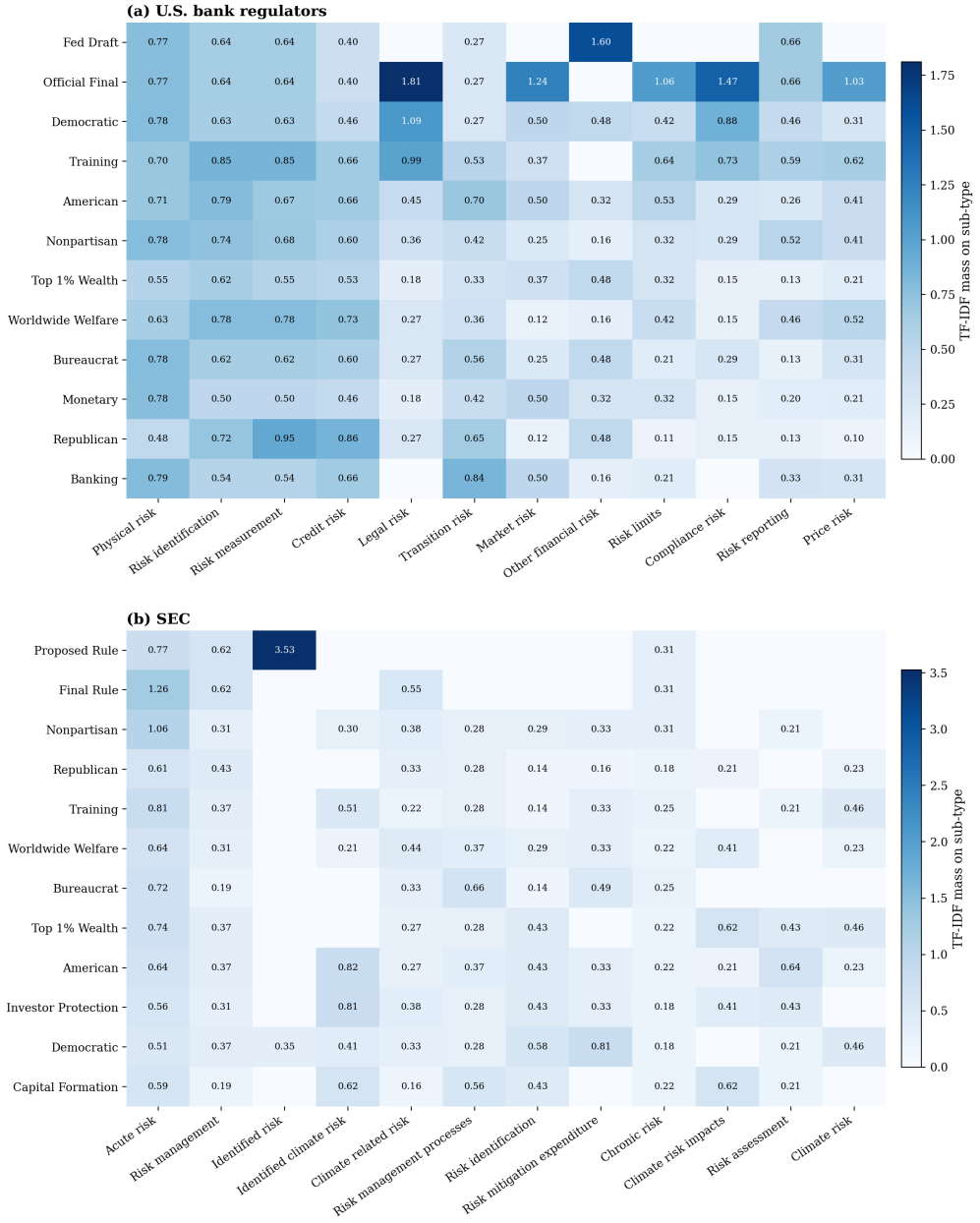


Figure OA7: Risk sub-types emphasized by each role-revised policy using GPT-5.2: (a) U.S. bank regulators and (b) the SEC.

Each cell reports the TF-IDF mass that the row's policy document places on the column's risk sub-type. Risk sub-types are all canonical topics whose phrase contains "risk," with case and plural variants summed, and the twelve sub-types with the most mass are shown. Topics are extracted by `gpt-5` from the GPT-5.2 role-revised policies. The first two rows are the draft and final policy: the Fed's draft and the final guidance in panel (a), and the proposed and final rule in panel (b). The remaining ten rows are role-revised policies averaged across the ten runs, ordered top-to-bottom by the cosine similarity of their risk sub-type weights to those of the final policy. Columns are ordered left-to-right by total mass across rows. Darker blue indicates higher TF-IDF mass. *Capital Formation* and *Investor Protection* are the SEC counterparts of *Monetary* and *Banking*.

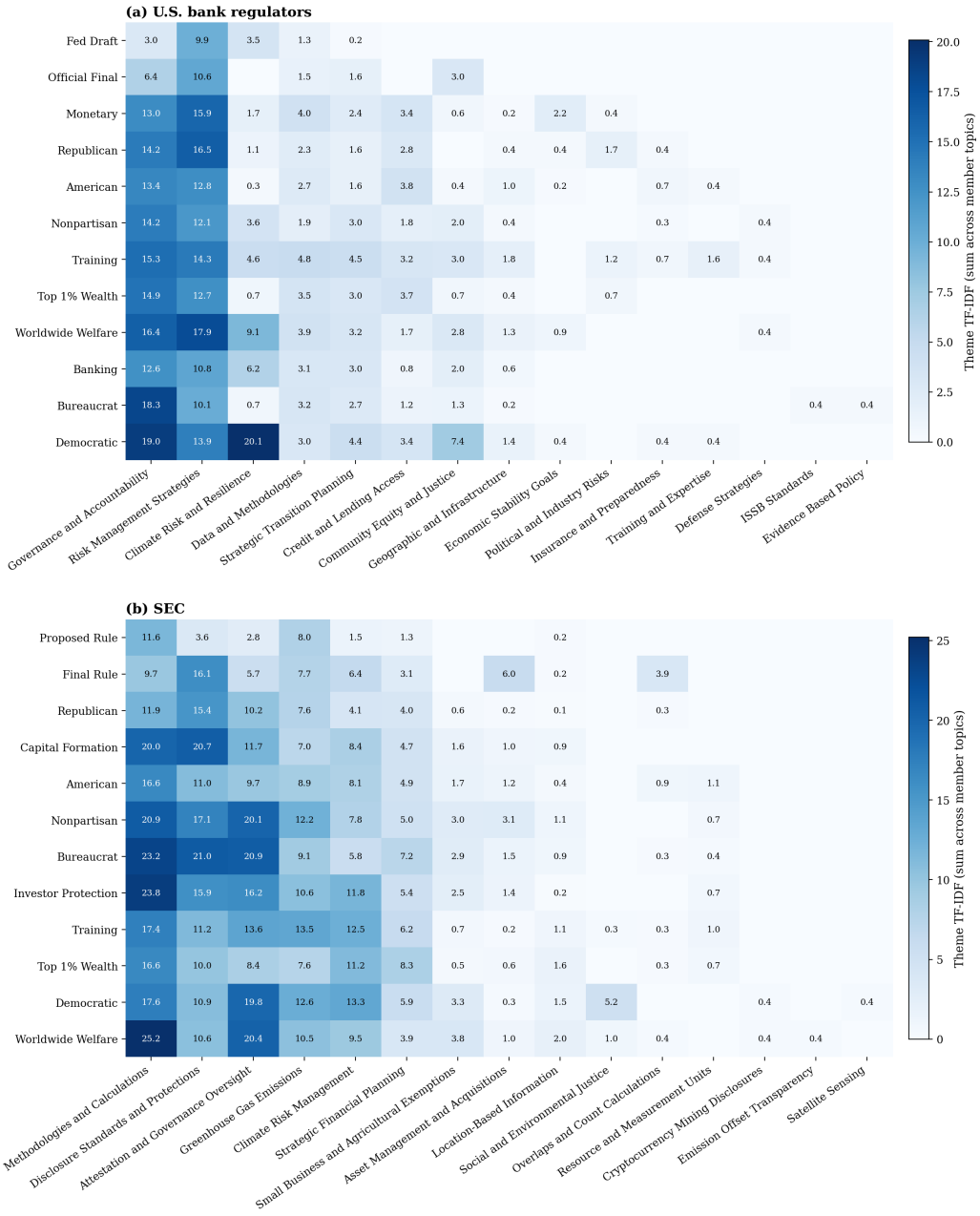


Figure OA8: Themed TF-IDF emphasis by policy document under Claude Sonnet 4.5: (a) U.S. bank regulators and (b) the SEC.

Each cell reports the TF-IDF mass that the row's policy document places on the column's theme, computed as the sum of TF-IDF scores of the topics assigned to that theme. The role-revised policies are produced by Claude Sonnet 4.5 under the same two-stage accept-and-revise pipeline used in Section 5.4 and Section 5.5. Topics are extracted by `gpt-5` and merged at a 0.90 embedding-cosine threshold, yielding 449 canonical topics for the bank regulators and 812 for the SEC. Themes are constructed separately for each regulator by clustering the topics' `text-embedding-3-large` embeddings into fifteen groups, which an LLM labels, so the themes differ across panels. The first two rows are the draft and final policy; the remaining ten rows are role-revised policies, each averaged across the ten runs. Rows are ordered top-to-bottom by themed-TF-IDF cosine similarity to the final policy (descending). Columns are ordered left-to-right by total mass across rows (descending). Darker blue indicates higher TF-IDF mass.

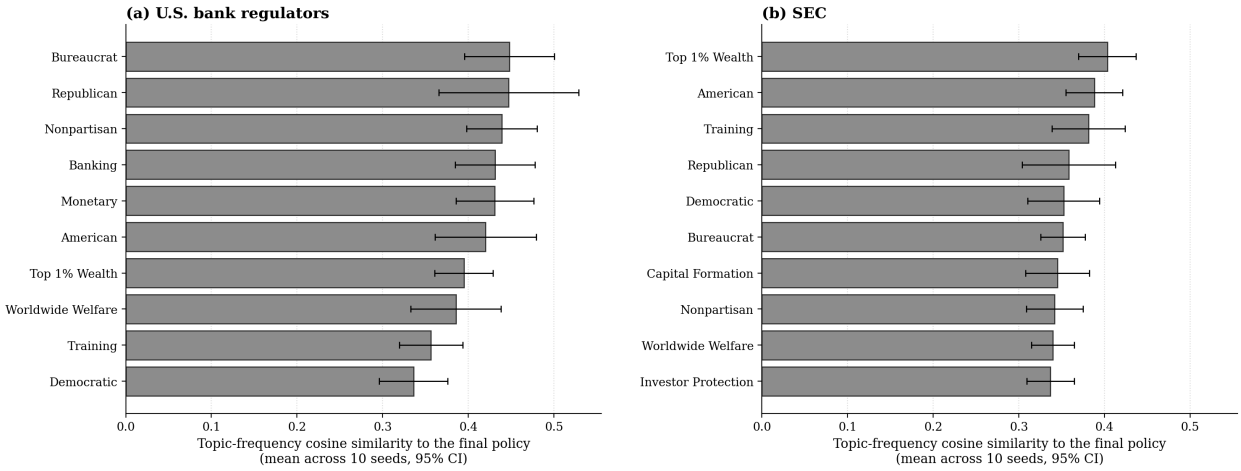


Figure OA9: Topic-frequency cosine similarity between role-revised policies and the final policy under Claude Sonnet 4.5: (a) U.S. bank regulators and (b) the SEC.

Horizontal bars show the mean cosine similarity between each role's revised policy and the final policy, computed on topic-frequency (TF) vectors over the canonical topics extracted by `gpt-5` from the Claude Sonnet 4.5 role-revised policies and merged at a 0.90 embedding-cosine threshold (449 topics for the bank regulators and 812 for the SEC). The TF vector for a document is a 0/1 indicator over the topic vocabulary. Means are taken across the ten runs; error bars are 95% confidence intervals, ± 1.96 standard errors, with the standard error computed from the standard deviation across runs. Roles are sorted top-to-bottom by mean similarity (descending). Both panels share the same horizontal axis.

F. Do the models remember the final policies?

This appendix describes the tests summarized in Section 5.7. We test the two commercial models that write the second-stage revisions in both settings, GPT-5.2 and Claude Sonnet 4.5, in the versions used in the main analysis. We set the temperature to zero, so that each model gives its most likely answer, and GPT-5.2 answers without reasoning, as in the main analysis. The tests use the texts of the main analysis: for the bank regulators, the final guidance and the Fed's draft; for the SEC, the codified text of the final and the proposed rule.

F.1. Completing paragraphs of the final policy

We use every paragraph of the final policy with at least 30 words: 19 paragraphs of the final guidance and 58 of the final rule. The model receives the first half of a paragraph and writes the next words, at most 50. Following Golchin & Surdeanu (2024), we complete each paragraph under two instructions. The guided instruction names the document, for example "the final Principles for Climate-Related Financial Risk Management for Large Financial Institutions, issued jointly by the Board of Governors of the Federal Reserve System, the Federal Deposit Insurance Corporation, and the Office of the Comptroller of the Currency, and published in the Federal Register on October 30, 2023," and asks the model to continue the paragraph, "reproducing the published wording as closely as you can from memory." The general instruction says only "Complete the following text with its next $[k]$ words." In a pilot, both models declined some guided requests as requests for copyrighted text, although works of the U.S. government are not copyrighted. The guided instruction therefore states that the document is a public-domain work of the U.S. federal government, and a system message explains that the request is part of a study of what language models remember. Even so, Sonnet declined 4 of the 58 SEC paragraphs under the guided instruction, and GPT-5.2 declined 2 under the general instruction. We compare the two instructions on the paragraphs that the model answered under both.

We compare each completion with the published continuation in three ways: the ROUGE-L score, which measures the longest common subsequence of words; the share of the continuation's eight-word sequences that the completion reproduces; and the longest run of consecutive words that the two share. Part of each final policy repeats its drafts, and a model that remembers a draft could reproduce that part without remembering the final policy. We therefore also report the share of the continuation's eight-word sequences that appear in none of the three drafts of the guidance, or nowhere in the SEC's proposing release (SEC, 2022c): the wording that first appeared in the final policy.

Panel A of Table OA10 reports the results. For the bank guidance, neither model reproduces a single eight-word sequence under either instruction, and no completion shares a run of more than five words with the published text. Naming the source raises GPT-5.2's ROUGE-L by 0.05 (95% bootstrap confidence interval 0.02 to 0.08) without producing any eight-word match, which suggests that the model writes more in the style of the guidance rather than that it recalls the text. For the SEC rule, the models reproduce 3.0–6.1% of the new eight-word sequences. Naming the source raises this share by 2.9 percentage points for GPT-5.2 (0.0 to 6.5) and changes it by −0.5 points for Sonnet (−4.1 to 2.9). Two of the 308 completions reproduce the published text exactly, and both are lists that the SEC took from earlier text. One is the list of disclosures about carbon offsets and renewable energy credits, which the proposal already contained and to which the final rule added the words "nature and"; the other is the list of examples of management's relevant expertise, which repeats the corresponding instruction of the SEC's 2023 cybersecurity disclosure rule (SEC, 2023) word for word apart from its subject. The longest runs of wording that is new relative to the proposal also come from that rule, whose definition of the board of directors of a foreign private issuer the final rule repeats. Sonnet also reproduces, under the guided instruction, the final rule's list of severe weather events, to which the final rule added hurricanes and tornadoes.

F.2. Writing the adopted version of each draft paragraph

We pair each paragraph of the final policy with the draft paragraph that it replaced: for the bank regulators, the matched paragraph of the Fed's draft (Online Appendix A; 19 pairs), and for the SEC, the paragraph of the proposed rule text whose embedding is most similar (78 pairs, with 42 distinct draft paragraphs). Under the guided instruction, the model is told the source of the draft paragraph and the name and date of the final policy, and is asked to write the adopted paragraph, "reproducing the adopted wording as closely as you can from memory," keeping the draft wording wherever it does not remember a change. Under the general instruction, it is asked to revise the paragraph "as the regulator might have revised it for the final version after considering the public comments." We count the four-word sequences of the adopted paragraph that are not in the draft paragraph, which is the wording the regulators added, and the four-word sequences of the draft paragraph that are not in the adopted one, which is the wording they removed. A model that remembers the adopted paragraph writes the added wording and drops the removed wording. A model that does not remember it, and follows the instruction, returns the draft, which writes none of the added wording and keeps all of the removed wording. We also compare the embedding of the model's paragraph with that of the adopted paragraph, next to the comparison of the draft with the adopted paragraph.

Panel B shows that, under the guided instruction, both models return the draft almost unchanged. They write at most 0.9% of the added sequences and keep at least 92% of the removed ones, and their paragraphs are no closer to the adopted paragraphs than the drafts are. Sonnet states that it does not remember the adopted wording for 51 of the 78 SEC paragraphs; for 43 of them it then returns the draft, as instructed, and for the other 8 it declines to write a paragraph, and we drop these 8. Under the general instruction, the models change more of the draft, but their changes match the regulators' changes no better: they write at most 2.7% of the added sequences.

F.3. The SEC's main decisions

We ask each model four multiple-choice questions about the final rule, each with three answers: the adopted provision, the proposed provision, and a third alternative, in random order. The questions ask which registrants must disclose Scope 3 emissions, which must disclose Scope 1 and Scope 2 emissions, whether a registrant must disclose whether any board member has expertise in climate-related risks, and when a registrant must disclose in its financial statements the expenditures and losses caused by severe weather events. We take the adopted and proposed provisions from the codified text of the two rules. Both models answer all four questions correctly (Panel C).

F.4. Do the revisions use this knowledge?

Finally, we check whether the second-stage revisions of the SEC's proposal adopt the decisions that the models know. We search the 100 revisions of each model (ten roles, ten runs each) for three features: any provision on Scope 3 emissions, which the proposal contains and the final rule does not; the final rule's threshold for disclosing the financial effects of severe weather events, which is one percent of the absolute value of income or loss before income tax, with an exception for amounts under $100,000 (and a corresponding threshold for capitalized costs, with an exception for amounts under $500,000); and any sentence that mentions both the board and expertise, which the proposal contains and the final rule does not. Every revision addresses Scope 3 emissions, and none uses the final rule's threshold (Panel C). A provision on the climate expertise of board members remains in 81 of GPT-5.2's revisions and 79 of Sonnet's. Whether it remains depends on the role: under Sonnet, the *Republican* role keeps it in one of ten runs, the *Top 1% Wealth* role in five, and the *Democratic*, *American*, *Investor Protection*, *Worldwide Welfare*, and *Training* roles in all ten.

These tests measure what the models reproduce or state when asked. They cannot rule out that knowledge of the final policies shapes a revision in ways that the tests do not detect. But the revisions do not adopt even the decisions that both models identify correctly.

Table OA10: Do the models remember the final policies?

This table reports the tests described in Online Appendix F for GPT-5.2 and Claude Sonnet 4.5 at temperature zero. Panel A: the model receives the first half of a paragraph of the final policy and writes the next words (at most 50), with the source named (guided instruction) or not (general instruction). ROUGE-L is the F-score of the longest common subsequence of words between the completion and the published continuation. 8-word sequences: the share of the continuation's eight-word sequences that the completion reproduces. New 8-word sequences: the same share for the sequences that appear in none of the drafts (the three drafts of the guidance, or the SEC's proposing release). Longest shared run: the median (maximum) across paragraphs of the longest run of consecutive words that the completion shares with the continuation. Paragraphs: paragraphs of at least 30 words that the model answered under both instructions. Panel B: the model receives a draft paragraph and writes the adopted version (source named) or a revision (source not named). Added (removed) 4-word sequences are the four-word sequences of the adopted (draft) paragraph that are not in the draft (adopted) paragraph; the columns report the share of the added sequences that the model writes and the share of the removed sequences that it keeps. Cosine similarity compares `text-embedding-3-large` embeddings of the model's paragraph, and of the draft paragraph, with the adopted paragraph. Pairs: Sonnet declined 8 of the 78 SEC pairs under the guided instruction. Panel C: the number of correct answers to four multiple-choice questions about the SEC's final rule, and the number of the 100 second-stage revisions of the SEC's proposal by each model (ten roles, ten runs each) that contain any provision on Scope 3 emissions, the final rule's threshold for disclosing the financial effects of severe weather events, or a sentence on the board's expertise. The first two rows show whether the proposed and the final rule contain each feature.

Panel A. Completing paragraphs of the final policy

Model	Policy	Paragraphs	Source named	ROUGE-L	8-word sequences (%)	New 8-word sequences (%)	Longest shared run
GPT-5.2	Bank guidance	19	Yes	0.27	0.0	0.0	3 (5)
			No	0.23	0.0	0.0	3 (5)
GPT-5.2	SEC rule	56	Yes	0.38	7.0	5.9	4.5 (20)
			No	0.33	4.7	3.0	4 (17)
Claude Sonnet 4.5	Bank guidance	19	Yes	0.25	0.0	0.0	3 (5)
			No	0.24	0.0	0.0	3 (4)
Claude Sonnet 4.5	SEC rule	54	Yes	0.41	8.3	5.5	5 (36)
			No	0.40	9.7	6.1	4.5 (42)

Panel B. Writing the adopted version of each draft paragraph

Model	Policy	Pairs	Source named	Added 4-word sequences written (%)	Removed 4-word sequences kept (%)	Cosine similarity to adopted: Model's version	Cosine similarity to adopted: Draft
GPT-5.2	Bank guidance	19	Yes	0.9	96.7	0.972	0.972
		19	No	2.7	77.6	0.962	0.972
GPT-5.2	SEC rule	78	Yes	0.1	99.8	0.713	0.713
		78	No	0.5	95.6	0.714	0.713
Claude Sonnet 4.5	Bank guidance	19	Yes	0.9	98.3	0.971	0.972
		19	No	1.8	71.0	0.962	0.972
Claude Sonnet 4.5	SEC rule	70	Yes	0.1	92.6	0.699	0.705
		78	No	0.8	67.5	0.703	0.713

Panel C. The SEC's main decisions

	Questions	Revisions of the proposal (of 100) that		
Model	answered correctly (of 4)	address Scope 3 emissions	use the final rule's threshold	keep the board-expertise item
Proposed rule		Yes	No	Yes
Final rule		No	Yes	No
GPT-5.2	4	100	0	81
Claude Sonnet 4.5	4	100	0	79