

How does the CSRD affect processing costs and what role does GenAI play?*

Victor Wagner

Stockholm School of Economics

June 26, 2026

Abstract

In a 2×2 experiment, I randomly assign participants to analyze firms' reports before and after the CSRD reports and to either have access to a custom GenAI chatbot or not. I find that the CSRD lowers both acquisition and integration costs. While CSRD reports are not read faster, participants' answers are more accurate, especially for the questions that require more processing. Access to the GenAI bot does not affect processing costs. Its effect on accuracy is null; the apparent harm from heavier use is selection: granting access leads users to substitute the bot for reading. This hurts accuracy, as the bot does not relieve the integration-heavy task but instead displaces the reading that allows a user to catch the AI's errors. A large data-collection project (128 students, 750,000+ data points across >2,000 firm-years) confirms the CSRD pattern at scale and identifies the textual features driving the results.

*This paper is part of my dissertation at LMU Munich. For helpful comments and suggestions, I thank my dissertation chair, Thorsten Sellhorn, as well as Max Müller, Kevin Bauer, Matthias Breuer, Patricia Breuer, Peter Fiechter, Joachim Gassen, Ann-Kristin Großkopf, Christian Leuz, Lisa Lüttke (discussant), Gaizka Ormazabal, Lazo Papadopoulos, Ivo Schedlinsky, Katharina Schmidt (discussant), Harm Schütt, and Caren Sureth-Sloane, as well as seminar and conference participants at the University of Amsterdam, Bocconi, Chicago, Erasmus Rotterdam, FS, IE, INSEAD, LSE, SSE, WHU, the 1st Accounting Research Camp on Transparency in Corporations and Markets, and the TRR 266 PhD brown bag. Julian Flaschke and Sebastian Peschel provided excellent research assistance. The paper's AEA RCT ID is AEARCTR-0016263. Approval to conduct the experiment was granted by the ethics commission of the LMU Munich School of Management, No. ETH-SOM-007; approval for monitoring the data collection was granted by the ethics commission of LMU Munich's Faculty of Medicine, No. 23-0435. I acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG): project ID 403041268 – TRR 266. **Preliminary working paper — please do not circulate or cite without permission.** All errors are my own.

How does the CSRD affect processing costs and what role does GenAI play?

Abstract

In a 2×2 experiment, I randomly assign participants to analyze firms' reports before and after the CSRD reports and to either have access to a custom GenAI chatbot or not. I find that the CSRD lowers both acquisition and integration costs. While CSRD reports are not read faster, participants' answers are more accurate, especially for the questions that require more processing. Access to the GenAI bot does not affect processing costs. Its effect on accuracy is null; the apparent harm from heavier use is selection: granting access leads users to substitute the bot for reading. This hurts accuracy, as the bot does not relieve the integration-heavy task but instead displaces the reading that allows a user to catch the AI's errors. A large data-collection project (128 students, 750,000+ data points across >2,000 firm-years) confirms the CSRD pattern at scale and identifies the textual features driving the results.

1 Introduction

The rate at which firms disclose sustainability information has massively increased in recent years, as society demands greater accountability and transparency amid growing environmental and social crises. This rapid increase, however, has left many concerned about information overload, i.e., the high costs of processing these disclosures arising from their low comparability and the wide discretion firms enjoy in preparing them ([Alliance for Corporate Transparency 2020](#); [PwC 2023](#)). Responding to these concerns, the European Union has introduced the Corporate Sustainability Reporting Directive (CSRD, Directive EU 2022/2464), a far-reaching mandate that prescribes what firms report and in what form so as to “make sustainability information easily accessible for users” (recital 25). At the same time, the very act of processing these reports — sifting through lengthy documents to identify relevant information and integrating it into decision-making — is itself being transformed by the advent of Generative Artificial Intelligence (GenAI), which potentially renders the regulation’s benefits in reducing processing costs irrelevant. In this paper, I examine whether the CSRD reduces the processing costs of sustainability disclosures and what role access to GenAI plays.

I rely on the disclosure processing cost framework of Blankespoor, deHaan, and Marinovic ([2020](#)), which decomposes the cost of using a disclosure into three stages: becoming *aware* that a piece of information exists, *acquiring* it from the document, i.e., the cognitive effort of locating and reading the relevant content, and *integrating* it, i.e., combining it with other information into a judgment or comparison. The study focuses on the latter two components throughout.

I focus on the CSRD as a mandate to standardize sustainability disclosure for two reasons. First, it is a prominent instance of “targeted transparency,” i.e., disclosure mandated to steer firm behavior toward societal goals, which has become an important tool in the current policy environment ([Weil, Graham, and Fung 2013](#)). Second, and more importantly, the CSRD drastically changes the way sustainability information is presented, providing a viable shock to observe the effects on processing. This is necessary since firms have been publishing sustainability information since the 2000s, but until the CSRD they did so with wide discretion over what to report and in what form; the resulting disclosures were widely criticized as selective, incomparable, and self-serving ([Cho et al. 2015](#); [Milne and Gray 2012](#)), and sustainability reporting as a whole started from a far lower baseline of standardization than financial reporting, while featuring a diverse range of topics, little to no monetization, and a broader audience than investors alone ([H. B. Christensen, Hail, and Leuz 2021](#);

[Wagenhofer 2024](#); [Dechow 2023](#)). Because the CSRD removes precisely this discretion, its effect on processing costs should be more visible here than in settings — such as the move to International Financial Reporting Standards (IFRS) — where a common reporting culture and an enforcement infrastructure were already in place (for a review, see, e.g., [De George, Li, and Shivakumar 2016](#)).

The second lever in my study is Generative AI (GenAI). GenAI is already transforming the way preparers and users of financial and sustainability information work: firms have begun drafting their disclosures with it ([Blankespoor, deHaan, and Li, n.d.](#)), while investors and accountants increasingly read, analyze, and act on those disclosures through it ([Choi and Xie 2026](#); [Blankespoor, Croom, and Grant 2026](#); [Ecker et al. 2026](#); [Bradshaw et al. 2026](#)). Ex ante, however, the effect of AI on processing is ambiguous. On the one hand, recent large-scale field experiments document sizable productivity gains for workers equipped with generative AI, ranging from software developers ([Cui et al. 2026](#)) to call-center agents handling more routine tasks ([Brynjolfsson, Li, and Raymond 2025](#)). On the other hand, these gains are confined to a subset of tasks along a “jagged technological frontier” ([Dell’Acqua et al. 2026](#)), and a recent meta-analysis finds that human–AI combinations, on average, do not perform better than either humans or AI alone, with the shortfall concentrated in tasks that require judgment rather than mere retrieval ([Vaccaro, Almaatouq, and Malone 2024](#)). It is an open question, therefore, whether access to AI lowers the processing costs of complex corporate disclosures, let alone whether it can substitute for the CSRD.

To answer these questions, I integrate an experiment into a course on GHG emissions reporting under the CSRD, designed for sustainability practitioners and accounting students. In a 2×2 design, I randomly allocate 436 participants to reports from before or after the CSRD and, independently, to having access to a custom GenAI-based chatbot, and I measure the time they take and the accuracy of their answers when assessing three firms’ emissions performance. To complement this evidence and gauge its generalizability, I leverage a large data collection project in which 128 collectors recorded over 750,000 emissions data points on 944 firms (2,164 firm-years) for pre-CSRD and CSRD reports. This second setting lets me validate the constructs used in the experiment, confirm that the effect of the CSRD holds outside the experiment, and pin down which report features drive processing costs.

My main finding is that the CSRD reduces processing costs. Participants in the experiment take no less time to analyze CSRD than pre-CSRD reports, yet those who work with CSRD reports answer about 8 percentage points more of the questions correctly ($p < 0.001$) — a 12% improvement on the pre-CSRD baseline of 64%. Given that firms’ environmental disclosures became substantially

longer after the CSRD, the regulation does not so much cut the time a user spends as raise what that time yields, i.e., the amount of information she correctly extracts per unit of effort. The benefit, moreover, is concentrated in the harder questions — above all the one that requires integrating information across firms rather than merely locating a single figure — whereas accuracy on the easiest lookup, if anything, declines slightly, indicating that the CSRD eases the more cognitively demanding steps of processing rather than the step that was never costly. My field evidence shows the same pattern one step earlier in the workflow: CSRD reports carry far more mandated data points, yet collectors extract each one in markedly less time, so that here too, the regulation raises the yield of processing rather than its speed — and it does so along precisely the dimensions, i.e., information dispersed across the report and written in complex language, that the ESRS regularize.

Access to AI, by contrast, does not reduce processing costs. The effect of being granted access is statistically indistinguishable from zero for time and accuracy, and it neither amplifies nor erodes the benefit of the CSRD. Simply comparing *users* of the GenAI tool to non-users even reveals a negative effect. This, however, is an artifact of selection: it is the struggling participants who turn to the tool, and the effect disappears once I instrument AI use with random assignment. My analysis of the logged prompts and responses shows that the chatbot returned the correct figure during the conversation in over 91% of cases. Yet, receiving correct answers did not translate into better performance. This is because access to the tool resulted in substitution: granted the bot, participants read the underlying report less and query the tool in its place, at no saving in total time. The bot lowers the cost of acquiring a figure, i.e., locating and reading it, but the harder questions require integrating that figure into a judgment across firms, which the tool cannot do for the user. Neither the form of a participant's prompts nor the correctness of the bot's answers predicts how well she performs. In addition, reading the report is also what catches the answers the bot gets wrong; so the participants who rely on it most are left with no fallback when it fails. If anything, access to AI is mildly detrimental, and the data hints, without the precision to confirm it, that it dampens the learning participants would otherwise gain from completing additional rounds.

My paper offers several contributions. First, I contribute to the literature on the standardization of corporate disclosures by examining how the CSRD changes the cost of processing a report, complementing work in financial reporting that documents the comparability benefits of common metrics and formats (Leuz and Wysocki 2016; De Franco, Kothari, and Verdi 2011; DeFond et al. 2011; Yip and Young 2012) as well as work on the introduction of IFRS (Brüggemann, Hitz, and Sellhorn 2013; De George, Li, and Shivakumar 2016). Second, my study introduces a measure of processing

costs at the individual user level, complementing more indirect measures in the processing-cost literature (Blankespoor, deHaan, and Marinovic 2020). Third, as my field settings allow me to study how real reports are processed, my study adds to the evolving literature on the use of generative AI in real-world contexts (Cui et al. 2026; Brynjolfsson, Li, and Raymond 2025; Dell’Acqua et al. 2026; Choi and Xie 2026).

Finally, my findings speak directly to standard-setters. As the European Commission debates its “Omnibus” initiative, which proposes to cut the number of mandatory data points and make more disclosures voluntary (EU 2025), my results show that the CSRD has already lowered the processing costs of sustainability disclosures — and that it has done so along exactly those dimensions that, in the prior voluntary regime, made these reports costly to process. Whether granting firms more reporting discretion is desirable is, against this backdrop, far from obvious, particularly as the investors and user groups who increasingly rely on these disclosures for their decisions call for the standardization to be retained (Eurosif et al. 2025).

2 Institutional background and conceptual framework

2.1 The CSRD and potential effects on processing costs

The CSRD (Directive EU 2022/2464) is the largest-ever introduction of mandatory sustainability reporting. Enacted in 2023 as the successor to the Non-Financial Reporting Directive (NFRD; Fiechter, Hitz, and Lehmann (2022)), it initially brought roughly 50,000 firms into scope, with the first cohort reporting under the new rules for fiscal year 2024. Its core consists of the twelve European Sustainability Reporting Standards (ESRS) — two cross-cutting standards on general requirements and general disclosures, plus ten topical standards spanning environmental, social, and governance matters — which a firm applies to whichever topics its “double materiality” assessment deems relevant, i.e., topics that are material either to the firm’s enterprise value or to its impact on people and the environment. Where a topic *is* material, the firm loses its former discretion to choose its own metrics and layout: it must report defined data points (under ESRS E1, gross Scope 1–3 greenhouse-gas emissions (E1.44), emissions intensities (E1.53), or reduction targets (E1.30)) computed with a prescribed methodology and terminology (E1.46-52 and E1.AR39), presented in a prescribed tabular format (E1.AR48), and placed in a prescribed location within the management report (ESRS 1.110).

Standardization is, in this sense, a multifaceted construct — it fixes which figures a firm re-

ports, how it computes them, how it lays them out, and where it places them in the report — and these dimensions are hard to disentangle. Evidence on the CSRD’s predecessor, the 2014 Non-Financial Reporting Directive (NFRD; Directive EU 2014/95), shows that such a mandate is far from cosmetic: its arrival moved market prices ([Grewal, Riedl, and Serafeim 2019](#)) and changed firm behavior, prompting firms in scope to expand their real CSR activities even before the rules took effect ([Fiechter, Hitz, and Lehmann 2022](#)).

That the introduction of such a reporting mandate should lower processing costs rests on a substantial body of work in financial reporting. There, uniform metrics reported in a common format make disclosures more comparable, and a large literature documents that such comparability lowers the cost of acquiring information from a report and integrating it across firms ([Leuz and Wysocki 2016](#); [De Franco, Kothari, and Verdi 2011](#); [Barth et al. 2012](#); [DeFond et al. 2011](#); [Yip and Young 2012](#)). Direct evidence on how users process disclosures points the same way: more readable filings draw stronger reactions from small investors, who lean on them more heavily when the information is easier to process ([Rennekamp 2012](#)), and individuals invest more in firms whose disclosures are clear and concise, since complexity raises the cost of acquiring the relevant content ([Lawrence 2013](#)).

Applied to sustainability disclosure, the same logic predicts that prescribing what firms report, and in which format, should cut acquisition costs above all, i.e., the effort of locating and reading the relevant figure, since a user no longer has to decipher each firm’s idiosyncratic presentation before extracting it. The evidence on data workers points the same way: [Daske, Seregini, and Uckert \(2023\)](#), who examine the standardization of accounting terminology, find that those who collect and populate financial databases face lower processing costs for firms whose disclosures speak a common language. Lowering the cost of acquiring a figure, however, need not change what a user does with it: even when earnings information is made readily available, individual investors’ trades fail to incorporate it, which locates the residual binding constraint not in acquiring information but in integrating it into a judgment ([Blankespoor et al. 2019](#)).

The benefits of standardization, however, need not reach every user equally. When the SEC mandated XBRL (eXtensible Business Reporting Language) structuring of financial statements, the structured data did lower acquisition costs, yet the gains accrued disproportionately to sophisticated users already equipped to exploit them, rather than to the less sophisticated investors the mandate was meant to help ([Blankespoor 2019](#); [Blankespoor, Miller, and White 2014](#); [Bhattacharya, Cho, and Kim 2018](#)). The comparability gains from IFRS adoption tell a similar story: they materialized

mainly for firms with strong compliance incentives and credible enforcement, not wherever the standard nominally applied (Cascino and Gassen 2015; Daske et al. 2013; Brüggemann, Hitz, and Sellhorn 2013). Standardization, in other words, lowers processing costs on paper, but how much it does so in practice depends on who is reading and under what conditions.

Whether these lessons carry over to sustainability disclosure is, moreover, far from obvious, because, given its unique characteristics (Wagenhofer 2024), it is harder to standardize than financial reporting. Double materiality leaves firms wide latitude over which topics to report at all; many of the underlying concepts are interpretive rather than quantitative; and much of the resulting text remains boilerplate (H. B. Christensen, Hail, and Leuz 2021). Studies find that more disclosure can even *increase* disagreement among users rather than resolve it (Berg, Kölbel, and Rigobon 2022; D. M. Christensen, Serafeim, and Sikochi 2022). In addition, the mandate may even work against itself: when reporting regimes expand and tighten their disclosure requirements, firms often respond with longer and more complex disclosures that can overwhelm users rather than aid them (Chychyla, Leone, and Minutti-Meza 2019; Dyer, Lang, and Stice-Lawrence 2017), and a broad literature links such complex, less readable language to a lower accessibility of the underlying information (Miller 2010; Guay, Samuels, and Taylor 2016). Whether the CSRD lowers processing costs is therefore an open, empirical question.

2.2 Access to generative AI

The rapid adoption of generative AI (GenAI) has attracted considerable scholarly attention. Within accounting, an emerging literature documents that the technology lowers the costs of producing, disseminating, and processing financial information across all three dimensions (Cao et al. 2026). On the production and dissemination side, intermediaries who adopt GenAI become markedly more productive: authors on a crowdsourced research platform publish about a third more articles and broaden their firm coverage once they adopt the tool, which speeds price discovery and improves liquidity even as each individual article becomes less informative (Bradshaw et al. 2026). This dependence shows through in reverse when the tool is taken away — the temporary ban of ChatGPT in Italy reduced both the number and the accuracy of affected analysts' forecasts and widened bid-ask spreads, and its outages in the U.S. measurably depressed trading volume (Bertomeu et al. 2025; Cheng, Lin, and Zhao 2025).

Regarding processing, the technology is an information intermediary that sits between the user and the report and could, in principle, lower both acquisition costs — by locating and extract-

ing figures on demand — and integration costs — by summarizing and comparing across firms (Blankespoor, deHaan, and Marinovic 2020; Croom 2026). This idea is supported by findings that retail investors increasingly rely on GenAI to acquire and integrate information more quickly and make sense of complex data (Blankespoor, Croom, and Grant 2026; Ecker et al. 2026). An earlier search-facilitating technology over financial reports, XBRL, was shown to improve exactly this kind of acquisition and integration — but only for the users who engaged with it, and nearly half did not (Hodge, Kennedy, and Maines 2004). Recent field and laboratory experiments also point to benefits of AI: granting knowledge workers access to generative AI has been found to raise their productivity, with the largest gains often accruing to the least experienced workers and the spread in performance narrowing (Brynjolfsson, Li, and Raymond 2025; Noy and Zhang 2023; Chang et al. 2026).

The prior literature, however, cautions that any such benefit is neither automatic nor evenly distributed across tasks. Where generative AI does raise productivity, the gains tend to fall along a “jagged technological frontier” — sizable on some tasks yet absent, or even negative, on others that look superficially similar, and often largest for the least experienced workers (Cui et al. 2026; Dell’Acqua et al. 2026). Consistent with this, a recent meta-analysis finds that human–AI combinations, on average, do not outperform the better of the human or the AI alone, with the shortfall concentrated in tasks that require judgment rather than mere retrieval (Vaccaro, Almaatouq, and Malone 2024). Whether processing a sustainability report falls on the favorable side of that frontier — and, in particular, whether AI can ease the integration that the harder questions demand — is therefore an empirical question.

3 Experimental evidence on the CSRD and GenAI’s effects on processing costs

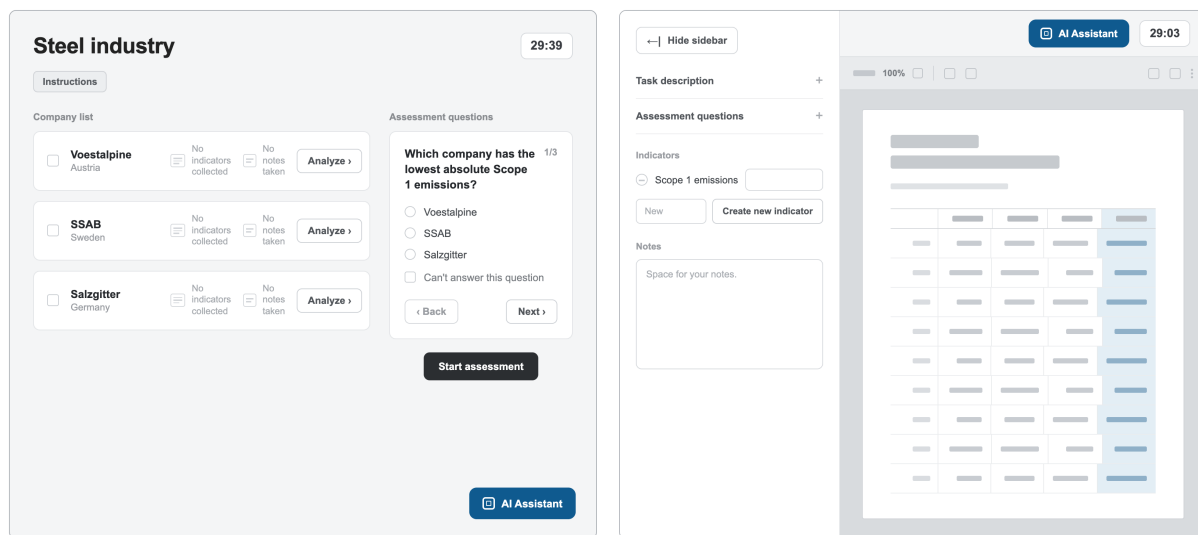
3.1 Participants, experimental setup, and measures

The core of the paper is an experiment I conducted as part of an introductory, three-hour online course on greenhouse gas (GHG) emissions reporting under the Corporate Sustainability Reporting Directive (CSRD). Specifically, participants have 30 minutes to analyze three real firms’ sustainability reports and then have to answer three questions about the firms’ GHG emissions performance. The course drew 436 participants. To get it started, I recruited 332 students from two sources: 205

from an introductory accounting course at my home institution, whom I incentivized through additional exam points, and 127 from other universities and recruitment channels, whom I incentivized with a monetary payment. The remaining 104 are non-student participants — predominantly practitioners, i.e., consultants, accountants, and sustainability reporting professionals — who enrolled in the course and completed the same assessment.

I run the experiment on a custom platform that I built to observe the different processing steps. After answering a short background questionnaire, each participant is randomly assigned an industry and shown a dashboard that lists the industry’s three firms on the left and the three assessment questions on the right. Opening a firm brings up its full report alongside a sidebar where the participant can define an indicator (e.g., absolute Scope-1 emissions in tonnes of CO₂-equivalents) and then record the value they found in the report. This arrangement serves three purposes at once: it lets participants carry their extracted figures into the assessment, since I close the reports while the questions are answered so that no one can simply look the answers up; it yields a clean measure of the time spent analyzing each report; and it captures, item by item, both what the participant extracted and whether that value was correct. Figure 1 shows a participant’s view at both stages.

Figure 1: The experimental platform



Notes. The platform as a participant sees it. *Left:* the industry dashboard, which lists the three firms to analyze, states the three assessment questions, and runs the 30-minute countdown; the AI assistant, when provisioned, is opened from the button at the lower right. *Right:* the analysis view of a single firm’s report, with a sidebar in which the participant defines indicators (e.g., absolute Scope-1 emissions), records the values found in the report, and takes notes. The reports are closed once the participant begins answering the questions, so the extracted figures must be carried over from this view.

I selected the industries based on the materiality of GHG emissions to their reporting and ensured

that all required information is present in the pre- and CSRD reports. Panel A, Table 1 lists the three industries and the respective firms. I drew the three questions from a review of how actual analysts assess emissions performance, choosing them to separate the two kinds of processing. Q1 and Q2 ask which of the three firms has the lowest absolute Scope-1 emissions and the lowest emissions intensity — the two headline indicators in this area, akin to a firm’s absolute profit and its profit margin. Each has a single disclosed figure behind it, so both are essentially acquisition tasks that a search tool should handle well. However, for pre-CSR D cases, emissions intensity was not a mandated measure, so participants might have to calculate it by hand, increasing acquisition and, potentially, integration costs. Q3 asks which firm has the most ambitious climate target, and has no single number behind it: answering it means weighing several disclosures at once — whether a net-zero commitment exists, how steep the targeted reduction is, against which baseline, and how far it leans on carbon offsets — which makes it an integration task that rests on the user’s own judgment.¹ The questions can be found in Panel B, Table 1.

Table 1: Experimental overview

Panel A: Industries and firms used		
Industry	Firms	
Airlines	Air France-KLM (France), Lufthansa (Germany), IAG (Spain)	
Cement	Cementir (Italy), Heidelberg Materials (Germany), Vicat (France)	
Steel	Salzgitter (Germany), SSAB (Sweden), Voestalpine (Austria)	
Panel B: The three outcome questions		
Question	What it requires	
Q1 Lowest absolute Scope 1	acquisition (lookup)	
Q2 Lowest emissions intensity	acquisition (lookup) and integration (calculation)	
Q3 Most ambitious climate target	integration (judgment)	
Panel C: 2×2 treatment design		
	Pre-CSR D reports	CSR D reports
Manual	baseline	CSR D effect
Access to AI chatbot	AI effect	both

Notes. **Panel A** lists the three industries and the firms shown in each. **Panel B** lists the three emissions questions and the type of processing each requires. **Panel C** presents the two independently randomized treatments: participants either see a firm’s pre-CSR D or CSR D report (columns) and have access to a custom AI chatbot over those reports (rows). A participant needs to successfully answer all three questions to complete the assessment and can finish all three industries to do so.

¹For Q3 to admit a defensible answer, the firms’ climate targets must be rankable along these dimensions. I therefore selected, within each industry, firms whose targets can be clearly ordered, and I confirmed the rankings with practicing ESG analysts before preregistering the study.

The two treatments were randomized independently within the participant pool: *CSRD*, which presents the same firms in their CSRD reports rather than their pre-CSRD reports, and *AI*, access to a custom chatbot I set up. The tool is built on a retrieval-augmented-generation (RAG) architecture: in contrast to an off-the-shelf chatbot, before answering, it retrieves the passages of the report most similar to the user’s query and supplies them to the underlying large language model, so that the answer rests on the firm’s actual disclosures rather than on the model’s training data — a common remedy for “hallucination”.² It thus pairs the generative, interactive interface of a chatbot with the firm-specific knowledge the task requires, and is, if anything, a more capable tool than what a typical user would assemble unaided.³ A participant needed to score all three questions correctly and could complete up to three industries, generating repeated rounds; the unit of observation is the participant-by-industry round. Panel C, Table 1 visualizes the treatment arms. Panel A, Table 2, gives an overview of the participants’ characteristics.

My design allows me to capture processing costs by two outcomes: *CORRECT*, the number of correct answers (0–3), and *TIME*, the minutes spent analyzing the reports. While I present results individually, I also aggregate both into a composite *SCORE*, where each component is expressed as a z-score relative to the sample, then winsorized and shifted to be non-negative. *ROUND* is the number of industries the participant is required to finish the assessment. Panel B of Table 2 provides means and standard deviations by group.

Due to the close administration of the experimental platform, I am also able to capture *QUERIES* (and *USED AI*) to see how much (or whether at all) a participant in the *AI*-treatment group used the chatbot, as well as measures relating to their processing behavior: *READING* captures the time spent in a firm’s report, not counting the time they opened the chatbot, which is captured separately as *BOT*, and *NAVIGATING* the time spent on the platform between the firm’s reports (e.g., reading the instructions and questions, finalizing notes). All measures are reported in minutes. Finally, I summarize each student’s prior background in a single *EXPERIENCE* factor, computed as the first principal component of the seven background questions (accounting courses taken, internships, having taken a sustainability course, program level, i.e., undergraduate or post graduate, and self-rated familiarity with sustainability reporting, report-reading frequency, and investment

²In their pioneering paper, Lewis et al. (2020) demonstrate that their Wikipedia-enriched RAG architecture answered Jeopardy questions more factually in 42.7% of cases than the benchmark language model without specific access to Wikipedia. This is because with RAG, the model retrieves relevant knowledge directly from the specified source (i.e., Wikipedia) before answering the question. In the sustainability domain, this architecture has been employed to extract information from sustainability reports at scale (Bronzini et al. 2024; Forster et al. 2026).

³The chatbot is not a mere research prototype: it shares its interface and RAG architecture with a tool fielded on a live sustainability-reporting platform, where users have submitted more than 30,000 queries.

experience). Panel C, Table 2 shows means and standard deviations for all variables, split by group.

Table 2: Descriptive statistics for the experiment

	Pre-CSR, no AI	Pre-CSR, AI	CSR, no AI	CSR, AI
Panel A: Participant characteristics				
<i>PARTICIPANTS</i>	116	124	100	96
<i>PRACTITIONER</i>	0.22	0.21	0.29	0.24
<i>N</i>	214	237	155	154
<i>FEMALE</i>	0.34 (0.47)	0.48 (0.50)	0.39 (0.49)	0.42 (0.50)
<i>AGE</i>	0.49 (0.79)	0.38 (0.73)	0.39 (0.67)	0.32 (0.59)
<i>EDUCATION</i>	2.34 (1.26)	2.17 (1.27)	2.12 (1.25)	2.10 (1.26)
Panel B: Experimental outcomes				
<i>SCORE</i>	2.13 (1.36)	2.16 (1.28)	2.48 (1.22)	2.36 (1.27)
<i>CORRECT</i>	1.88 (0.90)	1.93 (0.87)	2.15 (0.84)	2.08 (0.86)
<i>TIME</i>	19.48 (8.34)	19.72 (8.12)	19.01 (8.17)	19.54 (8.37)
<i>ROUND</i>	0.64 (0.73)	0.67 (0.77)	0.48 (0.69)	0.55 (0.69)
Panel C: Experimental interaction				
<i>USED AI</i>	—	0.73 (0.45)	—	0.72 (0.45)
<i>QUERIES</i>	—	5.97 (5.61)	—	6.05 (5.83)
<i>READING</i>	12.85 (9.41)	8.91 (6.82)	13.59 (8.61)	9.05 (7.08)
<i>BOT</i>	—	6.02 (6.18)	—	6.04 (6.32)
<i>NAVIGATION</i>	3.97 (5.96)	3.50 (4.78)	2.76 (3.97)	3.32 (5.07)

Notes. Means with standard deviations in parentheses beside them, by treatment arm, at the participant-by-industry round level. The four arms cross the CSR treatment (pre-CSR vs. CSR reports) with access to the AI chatbot; the two treatment indicators define the columns. *PARTICIPANTS* is the number of unique participants and *N* the number of participant-by-industry rounds in each arm; *PRACTITIONER* is the share of participants who are not students. **Panel A** reports participant characteristics, **Panel B** experimental outcomes, and **Panel C** how participants split each round between reading the report, querying the bot, and navigating. *USED AI*, *QUERIES*, and *BOT* are zero by construction in the arms without access to the chatbot (shown as a dash). All variables are defined in Appendix A.

The preferred specification estimates

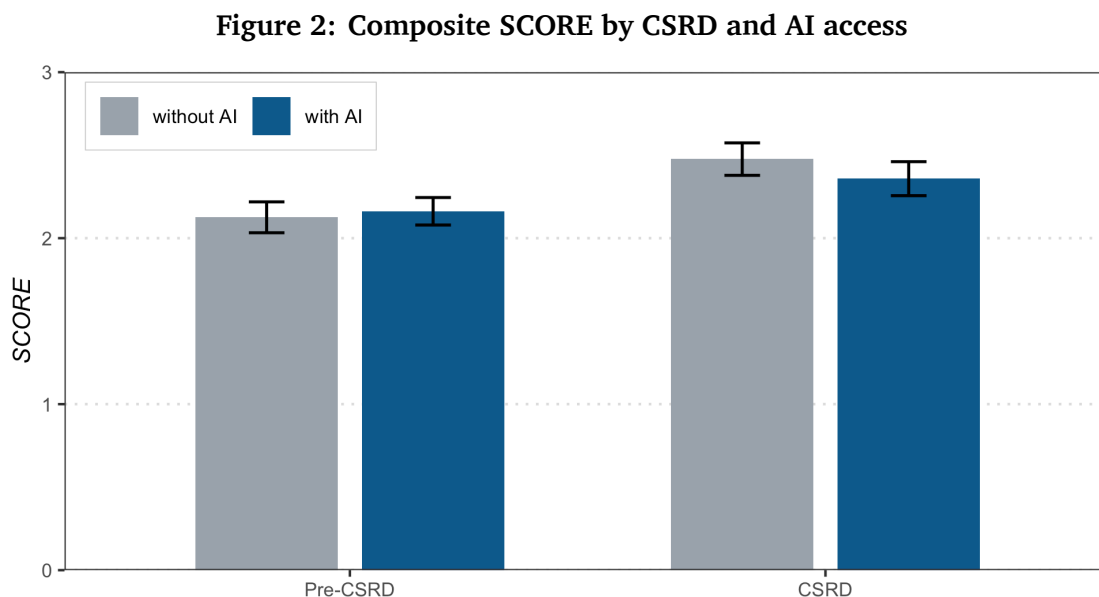
$$y_{ir} = \beta_1 CSR_i + \beta_2 AI_i + \beta_3 (CSR_i \times AI_i) + \gamma ROUND_{ir} + \alpha_{j(i,r)} + \varepsilon_{ir}, \quad (1)$$

where y_{ir} is the outcome of participant i in round r (*SCORE*, *CORRECT*, or *TIME*), AI_i and CSR_i are the randomized treatment indicators, $ROUND_{ir}$ is the order of the round, and $\alpha_{j(i,r)}$ is a fixed effect for the assigned industry; standard errors are clustered by participant. Because both treatments are randomly assigned, β_1 and β_2 identify the causal effects of the CSR and access to

AI, and β_3 — the direct test of whether the two levers substitute — identifies their interaction.

3.2 Main effects

I begin with the raw outcomes. Figure 2 plots the composite *SCORE* by arm, and the pattern the formal estimates will confirm is already visible: participants reading CSR D reports score markedly higher, while within each reporting regime, the arms with and without AI access are nearly indistinguishable.



Notes. Bar heights are the mean of *SCORE* in each treatment arm, crossing the CSR D treatment (pre-CSR D vs. CSR D reports) with access to AI (without vs. with AI); whiskers span ± 1 standard error. All variables are defined in Appendix A.

Table 2 confirms the composite pattern but illustrates the split in the outcomes: every arm spends roughly the same 19 minutes per round, but participants reading CSR D reports answer 2.12 of the three questions correctly against 1.91 with pre-CSR D reports.⁴

The regression-based estimates in Table 3 report the preferred specification for all three outcomes, thereby sharpening this picture. *CSR D* lifts both accuracy and the composite *SCORE*, while access to *AI* is indistinguishable from zero, and the two treatments do not interact. Participants working with CSR D reports answer about 8 percentage points more of the questions correctly (column 2; $p < 0.001$) — a 12% improvement on the pre-CSR D baseline of 64%, or roughly one additional correct answer for every 4 rounds completed — while taking no less time to do so (column 3). Given that firms' CSR D emissions sections are substantially longer than their pre-CSR D counterparts, the

⁴The fact that participants take roughly the same time is not surprising since the overall cap of 30 minutes was chosen to be rather strict.

standardized format more than makes up for the added reporting volume, i.e., the regulation raises the accuracy per unit of reading time rather than the speed of the task. The accuracy gain is stable as fixed effects and demographic covariates are progressively added, and it survives alternative completion-time windows.⁵

Table 3: Experimental outcomes

Dependent variable:	<i>SCORE</i>	<i>CORRECT</i>	<i>TIME</i>
	(1)	(2)	(3)
<i>CSRD</i>	0.399*** (0.134)	0.294*** (0.088)	-0.663 (0.926)
<i>AI</i>	0.013 (0.130)	0.036 (0.086)	0.270 (0.923)
<i>CSRD</i> × <i>AI</i>	-0.151 (0.202)	-0.110 (0.129)	0.346 (1.413)
<i>ROUND</i>	0.311*** (0.061)	0.155*** (0.041)	-1.204*** (0.391)
N	760	760	760
Adj. R^2	0.051	0.051	0.006
Industry FE	Yes	Yes	Yes

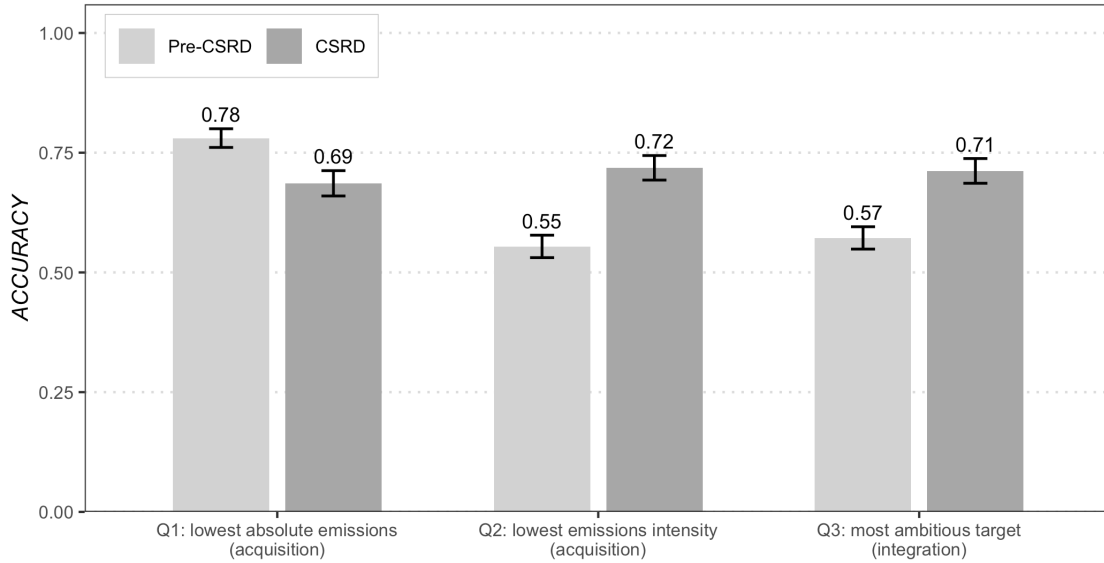
Notes. OLS estimates from the preferred specification. The dependent variable is *SCORE* in column 1, *CORRECT* in column 2, and *TIME* in column 3; each is regressed on *AI*, *CSRD*, their interaction *CSRD* × *AI*, and *ROUND*, with industry fixed effects; standard errors clustered by participant. All variables are defined in Appendix A.

The benefit of the CSRD, however, is not uniform across the three questions (Figure 3; Table 4). Accuracy on the first question, which asks about the lowest Scope-1 emissions declines ($p = 0.052$, column 1) for post-CSRD reports. Whereas 78% of the time, participants answered the question correctly with the shorter pre-CSRD reports, the question was answered correctly only 69% by participants with CSRD reports. One reason is that, even for such a headline figure, the sheer additional volume of reporting and with it, the introduction of more similar data points, makes it harder to access. Thus, the increase in accuracy is driven by Q2 ($p < 0.001$) and Q3 ($p < 0.001$), the questions that include integration. Both questions, Q2 with its requirement to calculate the intensity by hand and Q3 with its need to combine multiple dimensions of ambitiousness, were significantly more accurate for the groups reading CSRD reports. The CSRD, in other words, does

⁵The CSRD effect stays positive and the access-to-AI effect null across a specification ladder that adds industry fixed effects and participant demographics (gender, age, education, experience), and across completion-time windows of 3–40, 5–40, and 5–30 minutes.

not improve the step that was not costly (acquisition) but eases the steps where processing effort actually binds (integration).

Figure 3: Accuracy by question and CSRD treatment



Notes. Bar heights are the mean share of correct answers on each question (Q1, Q2, Q3), by CSRD treatment (pre-CSRD vs. CSRD reports); whiskers span ± 1 standard error. All variables are defined in Appendix A.

Table 4: Per-question treatment effects

Dependent variable:	<i>Q1 CORRECT</i>	<i>Q2 CORRECT</i>	<i>Q3 CORRECT</i>
	(1)	(2)	(3)
<i>CSRD</i>	-0.471* (0.242)	0.877*** (0.246)	0.855*** (0.242)
<i>AI</i>	-0.073 (0.230)	0.106 (0.193)	0.093 (0.200)
<i>CSRD</i> $\times$ <i>AI</i>	0.061 (0.346)	-0.218 (0.321)	-0.348 (0.330)
<i>ROUND</i>	0.443*** (0.120)	0.089 (0.098)	0.264** (0.109)
N	760	760	760
Adj. R^2	0.017	0.027	0.023
Industry FE	Yes	Yes	Yes

Notes. Logit estimates. The dependent variable is an indicator for *CORRECT* by answer on question Q1 (lowest absolute emissions) in column 1, Q2 (lowest emissions intensity) in column 2, and Q3 (most ambitious target) in column 3; each is regressed on *AI*, *CSRD*, their interaction *CSRD* $\times$ *AI*, and *ROUND*, with industry fixed effects; standard errors clustered by participant. All variables are defined in Appendix A.

Beyond the level effects, *ROUND* is the single largest driver of performance: participants improve markedly with each industry they analyze, and the improvement looks concentrated among the unassisted (Figure OA-1). The *ROUND* $\times$ *AI* interaction is negative — consistent with AI flattening the learning curve, i.e., with the skill decay and “cognitive debt” that reliance on an assistant has been argued to accumulate (Macnamara et al. 2024; Kosmyna et al. 2025) — but not significant at conventional levels (untabulated). With only three rounds per participant, the experiment has limited power to estimate these slopes, so I read the learning channel as suggestive rather than established.

3.3 GenAI access mechanism analysis

The null result regarding access to AI is informative, as the empirical design bounds the magnitude of any hidden benefit. The 95% confidence interval for the effect of access on the share of correct answers spans -4.5 to 3.9 percentage points⁶ — excluding, at the upper end, gains larger than about half the CSRD effect — and the interval for *SCORE* excludes improvements larger than 0.11 standard deviations. The average null result also does not mask a “leveling” effect of the kind documented for other knowledge-work tasks, where GenAI helps the least experienced most (Cui et al. 2026; Dell’Acqua et al. 2026): were that at work here, the effect of access should decline with the participant’s prior *EXPERIENCE*, but the interaction is small, statistically insignificant, and, if anything, positive ($p = 0.30$; untabulated). Access to AI thus provides no benefit on average and no leveling benefit for the least experienced.

However, having access to the tool did not automatically translate into use: 28% of rounds in the *AI* arm issued no query at all, so only about 72% of assigned rounds took the bot up, and among those that did, the median round issued just 8 queries. This partial take-up is why the effect of *access* — the intention-to-treat (ITT) effect, i.e., the effect of being offered the bot whether or not a participant goes on to use it — is the clean estimand of the design. Because the control arm cannot query the bot, random assignment also serves as a strong instrument for realized *use* (first-stage $F \approx 959$), recovering a local average treatment effect (LATE), i.e., the effect of using the bot for the participants who do so only because they were offered it. Any comparison conditioned on realized use, by contrast, sets self-selected users against non-users, and the results that follow must be read in that light.

Within the arm that could use the bot, more use predicts worse outcomes, not better: each

⁶The bounds are taken from the additive specification — access to *AI* and the *CSRD* entered without their interaction.

additional query is associated with a lower *SCORE*, fewer correct answers, and a longer round (Table 5). This is not the signature of a tool that helps those who lean on it, but of selection, i.e., the participants who struggle are the ones who turn to the bot most. This pattern thus describes who queries, not what querying does.⁷

Table 5: AI use and performance

Dependent variable:	<i>SCORE</i>	<i>CORRECT</i>	<i>TIME</i>
	(1)	(2)	(3)
<i>QUERIES</i>	-0.087*** (0.011)	-0.005 (0.008)	0.722*** (0.073)
<i>CSRD</i>	0.244* (0.141)	0.187** (0.094)	-0.263 (0.928)
<i>ROUND</i>	0.238*** (0.075)	0.147*** (0.056)	-0.682 (0.496)
N	391	391	391
Adj. R^2	0.180	0.051	0.248
Industry FE	Yes	Yes	Yes

Notes. OLS estimated within the AI-access arm only. The dependent variable is *SCORE* in column 1, *CORRECT* in column 2, and *TIME* in minutes in column 3; each is regressed on the number of AI queries issued in the round (*QUERIES*), *CSRD*, and *ROUND*, with industry fixed effects; standard errors clustered by participant. All variables are defined in Appendix A.

Having established what using the AI bot does, I now turn to why. I first examine how participants use it, i.e., the way access reshapes their effort. Here, the treatment arms visibly diverge. Table 6 decomposes each round into *READING* (time in the report), *BOT* (time querying the chatbot), and *NAVIGATION* (time moving between the two), first as arm means (Panel A) and then as the causal effect of access (Panel B). Granting access redistributes the way participants process reports. Participants given the bot spend about 4 fewer minutes on *READING* and about 6 more on *BOT*, while *NAVIGATION* and — tellingly — total time per round are unchanged ($p = 0.770$ for total time). They direct roughly 37% of their combined *READING* and *BOT* time at the chatbot, reading a 29% smaller slice of the disclosure than their unassisted counterparts. Access to AI thus causes participants to *substitute* the bot for the report rather than to consult it on top of the report.⁸

⁷The LATE confirms that this negative use-conditioned association is selection, not a causal penalty: the naive association between realized use and outcomes is negative, but instrumenting use with random assignment returns the estimate to essentially zero, in line with the ITT null on *access* (Online Appendix Table OA-1).

⁸The substitution is starkest, and least rewarded, when participants consult the bot from the overview screen without a single report open (about 15% of all queries). These rounds read the source for barely 3 minutes against 11 for the rest of the AI arm, letting the bot stand in for roughly three-quarters of their *READING*-plus-*BOT* time, and each such query

Standardization moves none of this allocation (Panel B, second row), consistent with its working through the content of the disclosure rather than through how users divide their attention.

Table 6: Time allocation per round

	<i>READING</i>	<i>BOT</i>	<i>NAVIGATION</i>	<i>TIME</i>
	(1)	(2)	(3)	(4)
Panel A: Mean minutes per round, by arm				
Pre-CSRSD, no AI	12.9	—	4.0	19.5
Pre-CSRSD, AI	8.9	6.0	3.5	19.7
CSRSD, no AI	13.6	—	2.8	19.0
CSRSD, AI	9.1	6.0	3.3	19.5
Panel B: Effect on time allocation (OLS)				
<i>AI</i>	-3.86*** (0.98)	6.01*** (0.53)	-0.49 (0.70)	0.27 (0.92)
<i>CSRSD</i>	0.42 (1.12)	0.01 (0.04)	-1.19* (0.68)	-0.66 (0.93)
<i>ROUND</i>	-1.96*** (0.39)	0.07 (0.21)	0.15 (0.23)	-1.20*** (0.39)
Industry FE	Yes	Yes	Yes	Yes
N	760	760	760	760
Adj. R^2	0.089	0.309	0.000	0.006

Notes. Each round is split into the minutes a participant spends reading the report (*READING*), querying the AI chatbot (*BOT*), and navigating (*NAVIGATION*); total time (*TIME*) is the full round, of which the three activities are the active portion (idle and answer-entry time are excluded, so they need not sum to it). **Panel A** reports arm means at the participant-by-industry round level. **Panel B** reports the effect of each treatment, from the same specification as the main models (access to *AI*, the *CSRSD*, their interaction, and *ROUND*, with industry fixed effects); each column is a separate regression of the corresponding time component, with standard errors clustered by participant; the chatbot is unavailable in the non-AI arms by construction. *, **, and *** denote significance at the 10%, 5%, and 1% levels. All variables are defined in Appendix A.

Whether this substitution harms performance depends on how good a substitute the bot is for reading. I thus analyze next the quality of the queries the participants ask the tool and the bot’s accuracy. To judge the quality of the queries, I used an LLM, first grouping each query into “keyword-only”, “partial query”, and “well-formed query” and then comparing the bots’ answers to the ground truth, i.e., the actually disclosed values. Only about 15% of the queries are “well-formed”, the rest are keyword-only, vague, or mis-scoped. Yet the quality of the prompts is not what limits carries a significantly lower *SCORE* (-0.12, $p < .01$) and a lower final accuracy. In its purest form, substituting the bot for the report actively hurts.

performance, because the bot answers well even when the prompts are imperfect. The bot returns the correct Scope-1 figure on 66% of individual lookups, and because participants ask repeatedly, it surfaces a correct figure at some point in about 91% of the rounds in which it is used. Retrieval, then, is not the weak link. More tellingly, the quality of the exchange does not predict performance at all: neither better-formed questions nor correct answers from the bot are associated with higher scores (Online Appendix Table OA-2, columns 1–3), whereas the *CSRD* and *ROUND* still are. What binds, similarly to the question-based analyses in the prior section, is not the acquisition of the information but integration, i.e., turning a correct number into the right cross-firm judgment. This is consistent with evidence that the difficulty with AI assistance lies less in what the model returns than in whether and how users incorporate it (Croom 2026; Bauer, Zahn, and Hinz 2023; Yang et al. 2026).

And precisely this is where the substitution documented in Table 6 turns costly. Because reading the report allows the user to catch a potential wrong answer of the bot when the participants stop reading, they give up that check. The bot’s first answer on a queried topic is wrong about 24% of the time, and what follows depends on whether the participant is still reading. When the bot’s first answer is correct, how much a participant reads barely moves accuracy (69% against 65% for the least- and most-reading halves) — the figure is already in hand.⁹ When that first answer is wrong, reading makes it still possible to answer the question correctly: accuracy climbs from 52% among those who read least to 74% among those who read most, as the readers catch the mistake while those who lean on the bot carry its error into their answer. Estimated directly, the interaction of reading time with a wrong first answer is positive and significant (logit coefficient 0.10, $p = 0.003$), whereas reading time on its own does not raise accuracy: its value shows up only once the bot has erred.¹⁰

To address the concern that the querying behavior documented here is an artifact of my experiment or the participant composition, I analyze a random sample of the logs of the platform where real users query the chatbot.¹¹ 6,431 judged queries from 2,140 users across 1,025 reports reveal that queries are about 15% bare keywords, 51% partial, and 34% well-formed, compared to the

⁹I split the queried topics within the *AI* arm at the median of *READING* time on the topic in question: those who read least are the below-median half and those who read most the above-median half, with accuracy measured on the same topic.

¹⁰This comparison is within the *AI* arm and observational — weaker participants tend both to read less and to score lower — so I read it as consistent with reading acting as a check on the bot, not as a clean causal estimate of that check. The direction is nonetheless telling: trading reading for the bot gives up the hedge that would otherwise catch the roughly one-in-four answers the bot gets wrong, leaving those who lean on it hardest without a net when it fails.

¹¹The chatbot shares its interface and RAG architecture with a tool that is deployed on a live sustainability-reporting platform where over 30,000 queries were submitted over a span of 12 months.

24/61/15% split in the experiment (Table 7). Thus, real users do prompt somewhat better — a composite weighted average that scores a keyword (0), a partial query (1), and a well-formed one (2) reaches 1.19 against 0.91 — but the gap is small set against how far both fall short of consistent well-formed querying: about two-thirds of real queries are still not well-formed, and roughly one in seven are bare keywords. The unstructured prompting I document in the experiment seems to be pervasive even for real users.

Nor is this poor prompting due to an unpractised sample. Splitting the judged conversations by whether the participant is a student or a practitioner shows that the two groups prompt near-identically: well-formed queries make up about 15% of practitioners’ queries against 14% of students’, a gap that a logit of a well-formed-query indicator on practitioner status, clustered by participant, leaves statistically insignificant ($p = 0.864$). Prior professional experience, thus, is not associated with better prompting.

Table 7: Prompt quality in the field versus the experiment

	Keyword	Partial	Well-formed	Weighted avg.
Experiment	0.24	0.61	0.15	0.91
Field platform	0.15	0.51	0.34	1.19

Notes. Shares of judged queries falling into each quality level — keyword, partial, or well-formed — for the experiment versus a 20% random subsample of the field platform’s real query log; rows sum to one. The weighted average scores a keyword 0, a partial query 1, and a well-formed query 2, so a higher value means better-formed prompting. All variables are defined in Appendix A.

Taken together, the experiment shows that while CSRD reports are read no faster, they are read more accurately, and the gain concentrates precisely where processing effort binds, i.e., in the intensity question where acquisition and integration are required, as well as in the integration-heavy target judgment. Access to GenAI has no effect and is thus no substitute for regulation: its effect on accuracy and on *SCORE* is statistically indistinguishable from zero, and the apparent harm from heavier use is selection rather than a causal penalty. The reason is mechanical. Granting access leads participants to substitute the bot for reading rather than to consult it alongside the report, and because the binding constraint is integrating a figure into the right cross-firm judgment rather than acquiring it, the tool relieves only the step that was already cheap. The reading it displaces, moreover, removes participants’ ability to spot one of the roughly one-in-four answers the bot gets wrong, so those who lean on it more often fail to answer the question.

4 Field evidence on the CSRD’s effect on processing costs

4.1 Empirical setup and measures

The random assignment of reading pre-CSRD versus CSRD reports in the experiment allows for causal identification of the CSRD’s effect. However, the limited sample of nine firms (three firms per industry) leaves two questions open: whether the documented positive CSRD effect generalizes beyond the controlled setting, i.e., is the result generalizable, and *which* disclosure characteristics drive the reduction in processing costs. The second part of the paper allows me to answer both questions by observing a large data-collection project: from March 2023 to January 2026, 128 collectors recorded more than 750,000 data points on 944 firms (2,164 firm-years) emission reporting for fiscal years 2020–2024. As part of their final theses, students were assigned a sample of firms; for every firm they first located and uploaded the annual and sustainability reports and then recorded 124 greenhouse-gas (GHG) emission-related data points, based on the three most relevant reporting standards ESRS E1, GRI 305, and IFRS S2.¹²

Similar to the first part, I set up the data collection environment as a website where students logged in to observe their information-processing behavior. The unit of observation is the collector–firm–year session, and the main outcome is again *TIME*, the minutes taken to collect one firm-year.¹³ Further measures come from the same close control of the platform. *DATA POINTS* is the number of data points found by the collector. Because each data point carries a timestamp, I observe not only total *TIME* but the collector’s accumulating experience, *SESSION*, an incrementing count of the firm-years she has already completed; and because collectors record the page on which they found each item, I can characterize the disclosure characteristics from the source pages themselves. Here, I construct a measure of how widely a firm-year’s data points are dispersed across the report (*DISPERSION*, a Herfindahl index of page numbers) as well as standard textual and layout properties of those pages (following [Lang and Stice-Lawrence 2015](#)): *DIFFICULT*, the Fog index, a measure for readability ([Li 2008](#)), the percentage of net positive (*TONE*), complex (*COMPLEX*), and uncertain (*UNCERTAIN*) words following Loughran and McDonald ([2011](#)), and the counts of tables and images.

¹²As the data collection started before the finalization of ESRS E1 and IFRS S2, I used the standard’s draft versions (November 2022) for the list of data points.

¹³To account for possible outliers due to inactivity, the *TIME* measure is aggregated from the *TIME* for the individual data points, where I exclude data point observations that took longer than 15 minutes.

Table 8: Descriptive statistics for the field observations

	Mean	SD	P25	Median	P75	N
Panel A: Data collection session characteristics						
<i>TIME</i>	33.42	14.36	22.39	31.47	42.96	3,741
<i>DATA POINTS</i>	21.52	16.40	9.00	19.00	30.00	3,741
<i>SESSION</i>	22.66	19.57	8.00	17.00	31.00	3,741
<i>CSRD</i>	0.07	0.26	0.00	0.00	0.00	3,741
Panel B: Disclosure characteristics						
<i>DISPERSION</i>	0.50	0.35	0.15	0.56	0.81	3,540
<i>DIFFICULT</i>	21.01	9.90	17.94	19.61	21.74	2,937
<i>TONE</i>	0.02	0.01	0.01	0.02	0.03	2,937
<i>COMPLEX</i>	0.00	0.00	0.00	0.00	0.00	2,937
<i>UNCERTAIN</i>	0.01	0.01	0.00	0.01	0.01	2,937
<i>TABLES</i>	7.14	6.98	2.40	5.57	9.60	2,942
<i>IMAGES</i>	3.99	53.84	0.00	0.83	3.00	2,942

Notes. Means, standard deviations, and quartiles, with *N*, for the field variables. The unit of observation is the collector-firm-year session. **Panel A** reports the data-collection session characteristics: *TIME* (minutes to collect one firm-year), *DATA POINTS* (the number recorded), *SESSION* (the count of firm-years the collector has already completed), and the *CSRD* indicator (one for the first mandatory CSRD reports, filed for fiscal year 2024). **Panel B** reports the source-page disclosure characteristics: *DISPERSION* (a Herfindahl index of the pages on which the data points are found), *DIFFICULT* (Gunning-Fog readability), *TONE*, *COMPLEX*, and *UNCERTAIN* (Loughran-McDonald dictionary shares), and the counts of *TABLES* and *IMAGES*; these are computed on the re-extracted source pages, an available subset, so their *N* is smaller. All variables are defined in Appendix A.

Panels A and B, Table 8 present descriptives on the data collection sessions and disclosure attributes. Across 3,741 sessions, it takes about 33 minutes to collect the 22 average data points a firm reports on emissions — a direct measure of the acquisition costs of disclosures. On average, the students collected data for about 29 firm-years, of which about 7% were published in accordance with the CSRD.

4.2 The CSRD's effect on processing costs

The specifications in Table 9 regress *TIME* on the *CSRD* indicator, allowing for a direct estimate of the CSRD's effect on acquisition costs. The columns add controls for the number of data points collected (log *DATA POINTS*) and internally generated experience (*SESSION*), and introduce firm fixed effects in columns (2) and (3). The results are in line with the experimental findings: given that CSRD reports contain more than twice as many mandated data points (49 versus 21), the total

time spent collecting data points is statistically indistinguishable from the pre-CSR D baseline once I condition on volume (Table 9, columns (1)–(2)). The *yield*, however, expressed in column (3), increases. What falls, sharply, is the cost *per data point*: from 2.8 to 1.1 minutes, or 55% lower once I hold the firm fixed ($p < 0.001$, column 3).¹⁴ To put the magnitude on a human scale: extracting the CSR D volume of roughly 49 data points at the pre-CSR D cost per item would take about 84 minutes longer per firm-year, i.e., well over an additional hour of an analyst’s attention for a single firm. As in the experiment, then, the CSR D raises the *yield* of acquisition — how much mandated information a collector extracts per minute — rather than the speed of the task as a whole.

Table 9: Acquisition cost pre- vs. CSR D

Dependent variable:	$\log(TIME)$	$\log(TIME)$	$\log(TIME)/DATA$ <i>POINTS</i>
	(1)	(2)	(3)
<i>CSR D</i>	-0.089 (0.145)	-0.040 (0.132)	-0.791*** (0.142)
$\log(DATA\ POINTS)$	0.250*** (0.016)	0.263*** (0.021)	
<i>SESSION</i>	-0.003** (0.001)	-0.004** (0.001)	-0.005*** (0.002)
N	3,542	3,542	3,542
Adj. R^2	0.256	0.317	0.486
Firm FE	No	Yes	Yes

Notes. OLS on emission-metric sessions. The dependent variable is the natural logarithm of total collection *TIME* in columns 1 and 2, and log minutes per data point in column 3. The *CSR D* indicator and *SESSION* enter every column; columns 1 and 2 also control for log *DATA POINTS* (the number of disclosed items), and columns 2 and 3 include firm fixed effects; standard errors are clustered by collector. All variables are defined in Appendix A.

Similar to the first part of the paper, the estimated coefficient on *SESSION* is negative and statistically significant in every specification (columns (1)–(3)): collection time falls as a collector completes more firm-years, with the gain steepest in the early sessions. This learning through internally generated experience mirrors the round-on-round improvement I document in the experiment, and it doubles as a construct-validity check — that effort visibly falls with practice confirms that *TIME*

¹⁴This per-data-point speed-up survives the sharpest possible matching. Restricting the sample to the 24,551 firm-by-data-point pairs (150 firms) that the same collectors recorded both before and under the CSR D, and absorbing a fixed effect for each firm-by-data-point cell, a disclosed item is located about 29% faster under the CSR D ($p = 0.042$). The same matching shows the regulation roughly doubles the share of these indicators that are actually disclosed (from 18% to 33%), and the speed-up is specific to disclosed items: the *CSR D* $\times$ *DATA POINTS* interaction is -0.40 ($p < 0.001$), whereas confirming that an item is *absent* is, if anything, slightly slower under the denser CSR D reports. Table OA-3 reports these estimates in the Online Appendix.

captures genuine processing effort.

4.3 Disclosure characteristics and processing costs

Having documented that the CSRD’s effect on processing cost reappears at scale, I now turn to identifying *which* disclosure characteristics make it costly to process, and whether they are the ones the CSRD regularizes. Table 10 regresses *TIME* on a set of reporting characteristics computed from the source pages, entered in three thematic groups — location, textual difficulty, and layout (columns (1)–(3)) — and then jointly (column (4)), within firm and year and controlling throughout for the collector’s accumulating experience (*SESSION*) and the number of data points collected (log *DATA POINTS*).

The cost drivers can be decomposed by theme rather than bundled together (Table 10), and the decomposition is sharp about which features bite and which do not. *DISPERSION* raises extraction time robustly, both on its own (column (1)) and with the other measures held fixed (column (4)). Among the textual measures (column (2)), it is the textual difficulty that carries through: *DIFFICULT* remains a robust driver in the joint specification (column (4)), whereas *COMPLEX* and *UNCERTAIN* do not survive alongside it, suggesting they proxy for the same textual difficulty that *DIFFICULT* already captures. *TONE* carries no weight. Among the layout measures (column (3)), pages that present the information in tables *reduce* extraction time (*TABLES*) whereas the number of *IMAGES* does not affect processing time.

The pattern maps almost one-to-one onto what the ESRS regularize: a prescribed location within the report (against dispersion), a common terminology and methodology (against dense, complex prose), and a prescribed tabular format for the emissions disclosures — the one page feature that demonstrably speeds extraction. The three-firm experiment cannot resolve this firm by firm; the field is where the *why* of the CSRD effect becomes visible, since the regulation regularizes precisely the dimensions that, in the voluntary regime, made reports costly to process.

Table 10: Reporting characteristics and acquisition time (DV = log time)

Dependent variable:	<i>log TIME</i>	<i>log TIME</i>	<i>log TIME</i>	<i>log TIME</i>
	(1)	(2)	(3)	(4)
<i>DISPERSION</i>	0.203*** (0.036)			0.194*** (0.037)
<i>DIFFICULT</i>		0.006*** (0.002)		0.006*** (0.002)
<i>TONE</i>		-1.506 (1.045)		-0.940 (1.010)
<i>COMPLEX</i>		9.699* (5.389)		6.860 (5.365)
<i>UNCERTAIN</i>		3.320* (1.949)		2.269 (1.859)
<i>TABLES</i>			-0.005*** (0.002)	-0.003* (0.002)
<i>IMAGES</i>			0.000 (0.000)	0.000 (0.000)
N	3,535	2,932	2,937	2,932
Adj. R^2	0.345	0.307	0.305	0.324
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Notes. OLS of log collection *TIME* on the reporting characteristics of the source pages, entered in three thematic groups: location (*DISPERSION*) in column 1; textual difficulty (*DIFFICULT*, *TONE*, *COMPLEX*, and *UNCERTAIN*) in column 2; and layout (*TABLES* and *IMAGES*) in column 3 — and all characteristics jointly in column 4. Every column also controls for *SESSION* and log *DATA POINTS* and includes firm and year fixed effects; standard errors are clustered by collector. All variables are defined in Appendix A.

Taken together, the field evidence generalizes the findings from the first part and extends the insights on the mechanisms: First, it shows that the CSRD effect generalizes far beyond nine firms: across 944 firms and 2,164 firm-years, CSRD reports carry more than twice as many mandated data points yet give up each one in markedly less time, so that here too the regulation raises the *yield* of acquisition rather than its speed. Second, it reveals *why*. The features that make a report costly to process — information dispersed across many pages and written in dense, difficult prose — are precisely the dimensions the ESRS regularize, through a prescribed location, a common terminology, and a tabular format that, of all the page features, is the one that demonstrably speeds extraction.

5 Concluding discussion

In this paper, I ask whether the Corporate Sustainability Reporting Directive (CSRD), the most far-reaching sustainability reporting mandate so far, lowers the cost of processing sustainability disclosures and what role access to generative AI plays. I study the questions with two complementary designs. I set up a randomized experiment as part of a course on emissions reporting, isolating the causal effect of the CSRD and of access to a retrieval-augmented (RAG) chatbot. A large field data-collection project, in which trained students extract data points from real reports, lends generalizability to the CSRD pattern and reveals which report features drive processing costs.

Of the two levers, only the regulation works. I find that the CSRD reliably lowers processing costs — not by saving time, but by raising the yield of a given amount of effort: CSRD reports are read no faster, yet they are read more accurately in the experiment and give up more extracted data points per minute in the field, supporting (Blankespoor 2019; DeFond et al. 2011; H. B. Christensen, Hail, and Leuz 2021; Krueger et al. 2024). Access to AI, by contrast, does not lower processing costs, and it is no substitute for the regulation: the apparent harm from the tool is selection, and the true effect of access is null. Because the binding constraint is integration rather than retrieval, even an accurate, well-explaining tool fails to help. Taken together, these results qualify the prevailing view that AI augments knowledge work for the specific task of processing regulated disclosures, placing the result on the unfavorable side of the human–AI frontier (Vaccaro, Almaatouq, and Malone 2024; Dell’Acqua et al. 2026).

I acknowledge several limitations of my study that point to future work. First, both empirical setups deal with real firms’ sustainability reports. While this increases the generalizability of the findings, it makes it harder to pinpoint which aspects of the CSRD drive the reduction in processing costs. In addition, I focus on emissions reporting, arguably the most standardized topic of sustainability reporting. With more granular measures and additional topics, future work can better dissect the drivers of standardization. Second, the participants in both designs are student-heavy. While I find no discernible differences in several analyses, future research could take inspiration from recent attempts to study “processing in the wild” (Blankespoor, Croom, and Grant 2026; Ecker et al. 2026). Third, my AI treatment is a single RAG-based design, so my evidence concerns *access* to this tool on this task, not a claim that no AI design could help; decomposing AI use and testing designs that support integration rather than mere lookup — including different degrees of interactivity (Croom 2026) — is the most promising avenue I see.

References

- Alliance for Corporate Transparency. 2020. “2019 Research Report: An Analysis of the Sustainability Reports of 1000 Companies Pursuant to the EU Non-Financial Reporting Directive.” Alliance for Corporate Transparency. <https://www.allianceforcorporatetransparency.org/>.
- Barth, Mary E., Wayne R. Landsman, Mark Lang, and Christopher Williams. 2012. “Are IFRS-Based and US GAAP-Based Accounting Amounts Comparable?” *Journal of Accounting and Economics* 54: 68–93. <https://doi.org/10.1016/j.jacceco.2012.03.001>.
- Bauer, Kevin, Moritz von Zahn, and Oliver Hinz. 2023. “Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users’ Information Processing.” *Information Systems Research* 34 (4): 1582–602. <https://doi.org/10.1287/isre.2023.1199>.
- Berg, Florian, Julian F Kölbel, and Roberto Rigobon. 2022. “Aggregate Confusion: The Divergence of ESG Ratings.” *Review of Finance* 26 (6): 1315–44. <https://doi.org/10.1093/rof/rfac033>.
- Bertomeu, Jeremy, Yupeng Lin, Yibin Liu, and Zhenghui Ni. 2025. “The Impact of Generative AI on Information Processing: Evidence from the Ban of ChatGPT in Italy.” *Journal of Accounting and Economics* 80 (1): 101782. <https://doi.org/10.1016/j.jacceco.2025.101782>.
- Bhattacharya, Nilabhra, Young Jun Cho, and Jae B Kim. 2018. “Leveling the Playing Field Between Large and Small Institutions: Evidence from the SEC’s XBRL Mandate.” *The Accounting Review* 93 (5): 51–71. <https://doi.org/10.2308/accr-52000>.
- Blankespoor, Elizabeth. 2019. “The Impact of Information Processing Costs on Firm Disclosure Choice: Evidence from the XBRL Mandate.” *Journal of Accounting Research* 57 (4): 919–67. <https://doi.org/10.1111/1475-679x.12268>.
- Blankespoor, Elizabeth, Joe Croom, and Stephanie M. Grant. 2026. “Generative AI and Investor Processing of Financial Information.” <https://doi.org/10.2139/ssrn.5053905>.
- Blankespoor, Elizabeth, Ed deHaan, and Qianqian Li. n.d. “Generative AI in Financial Reporting.”
- Blankespoor, Elizabeth, Ed deHaan, and Iván Marinovic. 2020. “Disclosure Processing Costs, Investors’ Information Choice, and Equity Market Outcomes: A Review.” *Journal of Accounting and Economics* 70 (2-3): 101344. <https://doi.org/10.1016/j.jacceco.2020.101344>.
- Blankespoor, Elizabeth, Ed deHaan, John Wertz, and Christina Zhu. 2019. “Why Do Individual Investors Disregard Accounting Information? The Roles of Information Awareness and Acquisition Costs.” *Journal of Accounting Research* 57 (1): 53–84. <https://doi.org/10.1111/1475-679x.12248>.

- Blankespoor, Elizabeth, Brian P. Miller, and Hal D. White. 2014. "Initial Evidence on the Market Impact of the XBRL Mandate." *Review of Accounting Studies* 19 (4): 1468–1503. <https://doi.org/10.1007/s11142-013-9273-4>.
- Bradshaw, Mark T., Chenyang Ma, Benjamin P. Yost, and Yuan Zou. 2026. "Generative AI Use by Capital Market Information Intermediaries: Evidence from Seeking Alpha." *Journal of Accounting Research* 64 (3): 1233–86. <https://doi.org/10.1111/1475-679x.70053>.
- Bronzini, Marco, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. "Glitter or Gold? Deriving Structured Insights from Sustainability Reports via Large Language Models." *EPJ Data Science* 13 (1). <https://doi.org/10.1140/epjds/s13688-024-00481-2>.
- Brüggemann, Ulf, Jörg-Markus Hitz, and Thorsten Sellhorn. 2013. "Intended and Unintended Consequences of Mandatory IFRS Adoption: A Review of Extant Evidence and Suggestions for Future Research." *European Accounting Review* 22 (1): 1–37. <https://doi.org/10.1080/09638180.2012.718487>.
- Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at Work." *The Quarterly Journal of Economics* 140 (2): 889–942. <https://doi.org/10.1093/qje/qjae044>.
- Cao, Sean Shun, Wilbur Xinyuan Chen, Guang Ma, and Suraj Srinivasan. 2026. "Generative AI in Capital Markets: Information Production, Dissemination, and Processing." *Journal of Accounting Research* 64 (3): 1427–50. <https://doi.org/10.1111/1475-679x.70061>.
- Cascino, Stefano, and Joachim Gassen. 2015. "What Drives the Comparability Effect of Mandatory IFRS Adoption?" *Review of Accounting Studies* 20 (1): 242–82. <https://doi.org/10.1007/s11142-014-9296-5>.
- Chang, Anne Yanru, Xi Dong, Xiumin Martin, and Changyun Zhou. 2026. "AI Democratization and Trading Inequality." *Journal of Accounting Research* 64 (3): 1287–1331. <https://doi.org/10.1111/1475-679x.70063>.
- Cheng, Qiang, Pengkai Lin, and Yue Zhao. 2025. "Does Generative AI Facilitate Investor Trading? Early Evidence from ChatGPT Outages." *Journal of Accounting and Economics* 80 (2-3): 101821. <https://doi.org/10.1016/j.jacceco.2025.101821>.
- Cho, Charles H., Giovanna Michelon, Dennis M. Patten, and Robin W. Roberts. 2015. "CSR Disclosure: The More Things Change...?" *Accounting, Auditing & Accountability Journal* 28 (1): 14–35. <https://doi.org/10.1108/aaaj-12-2013-1549>.
- Choi, Jung Ho, and Chloe L. Xie. 2026. "Human + AI in Accounting: Early Evidence from the Field." *Journal of Accounting Research* 64 (3): 1333–73. <https://doi.org/10.1111/1475-679x.70052>.

- Christensen, Dane M., George Serafeim, and Anywhere Sikochi. 2022. "Why Is Corporate Virtue in the Eye of the Beholder? The Case of ESG Ratings." *The Accounting Review* 97 (1): 147–75. <https://doi.org/10.2308/tar-2019-0506>.
- Christensen, Hans B., Luzi Hail, and Christian Leuz. 2021. "Mandatory CSR and Sustainability Reporting: Economic Analysis and Literature Review." *Review of Accounting Studies*. <https://doi.org/10.1007/s11142-021-09609-5>.
- Chychyla, Roman, Andrew J. Leone, and Miguel Minutti-Meza. 2019. "Complexity of Financial Reporting Standards and Accounting Expertise." *Journal of Accounting and Economics* 67 (1): 226–53. <https://doi.org/10.1016/j.jacceco.2018.09.005>.
- Croom, Joe. 2026. "Interactivity and Illusions of Ability: How Using Generative AI Affects Investor Judgments." *Journal of Accounting Research* 64 (2): 681–719. <https://doi.org/10.1111/1475-679x.70017>.
- Cui, Kevin Zheyuan, Mert Demirel, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. 2026. "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers." *Management Science*. <https://doi.org/10.1287/mnsc.2025.00535>.
- Daske, Holger, Luzi Hail, Christian Leuz, and Rodrigo Verdi. 2013. "Adopting a Label: Heterogeneity in the Economic Consequences Around IAS/IFRS Adoptions." *Journal of Accounting Research*. <https://doi.org/10.1111/1475-679x.12005>.
- Daske, Holger, Carol Seregini, and Matthias Uckert. 2023. "Unstandardized Accounting Terminology."
- De Franco, Gus, S. P. Kothari, and Rodrigo S. Verdi. 2011. "The Benefits of Financial Statement Comparability." *Journal of Accounting Research* 49 (4): 895–931. <https://doi.org/10.1111/j.1475-679x.2011.00415.x>.
- De George, Emmanuel, Xi Li, and Lakshmanan Shivakumar. 2016. "A Review of the IFRS Adoption Literature." *Review of Accounting Studies* 21 (3): 898–1004. <https://doi.org/10.1007/s11142-016-9363-1>.
- Dechow, Patricia M. 2023. "Understanding the Sustainability Reporting Landscape and Research Opportunities in Accounting." *The Accounting Review* 98 (5): 481–93. <https://doi.org/10.2308/tar-2023-0145>.
- DeFond, Mark, Xuesong Hu, Mingyi Hung, and Siqu Li. 2011. "The Impact of Mandatory IFRS Adoption on Foreign Mutual Fund Ownership: The Role of Comparability." *Journal of Accounting and Economics* 51 (3): 240–58. <https://doi.org/10.1016/j.jacceco.2011.02.001>.

- Dell’Acqua, Fabrizio, Edward McFowland, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Kraye, François Candelon, and Karim R. Lakhani. 2026. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality.” *Organization Science* 37 (2): 403–23. <https://doi.org/10.1287/orsc.2025.21838>.
- Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence. 2017. “The Evolution of 10-k Textual Disclosure: Evidence from Latent Dirichlet Allocation.” *Journal of Accounting and Economics* 64 (2-3): 221–45. <https://doi.org/10.1016/j.jacceco.2017.07.002>.
- Ecker, Frank, Xitong Li, Yilan Li, and Fan Wu. 2026. “How Stock Market Participants Use Generative Artificial Intelligence: Evidence from User-Platform Interaction Data.” *Journal of Accounting Research* 64 (3): 1375–1426. <https://doi.org/10.2139/ssrn.5224596>.
- EU. 2025. “Council and Parliament Strike a Deal to Simplify Sustainability Reporting and Due Diligence Requirements and Boost EU Competitiveness.” 2025. <https://www.consilium.europa.eu/en/press/press-releases/2025/12/09/council-and-parliament-strike-a-deal-to-simplify-sustainability-reporting-and-due-diligence-requirements-and-boost-eu-competitiveness/>.
- Eurosif, IIGCC, PRI, Corporate Leaders Group, E3G, and GRI. 2025. “Weakening the EU’s Sustainability Rules Risks Damaging Competitiveness and Growth, Warn Companies and Investors.” https://www.eurosif.org/wp-content/uploads/2025/07/2025.07.01-Investor-and-Business_Joint-Statement_PR.pdf.
- Fiechter, Peter, Jörg-Markus Hitz, and Nico Lehmann. 2022. “Real Effects of a Widespread CSR Reporting Mandate: Evidence from the European Union’s CSR Directive.” *Journal of Accounting Research* 60 (4): 1499–549. <https://doi.org/10.1111/1475-679x.12424>.
- Forster, Kerstin, Lucas Keil, Victor Wagner, Maximilian A. Müller, Thorsten Sellhorn, and Stefan Feuerriegel. 2026. “Assessing Corporate Sustainability with Large Language Models: Evidence from Europe.” *Nature Communications*.
- Grewal, Jody, Edward J. Riedl, and George Serafeim. 2019. “Market Reaction to Mandatory Non-financial Disclosure.” *Management Science* 65 (7): 3061–84. <https://doi.org/10.1287/mnsc.2018.3099>.
- Guay, Wayne, Delphine Samuels, and Daniel Taylor. 2016. “Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure.” *Journal of Accounting and Economics* 62 (2-3): 234–69. <https://doi.org/10.1016/j.jacceco.2016.09.001>.
- Hodge, Frank D., Jane Jollineau Kennedy, and Laureen A. Maines. 2004. “Does Search-Facilitating

- Technology Improve the Transparency of Financial Reporting?” *The Accounting Review* 79 (3).
- Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. “Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task.”
- Krueger, Philipp, Zacharias Sautner, Dragon Yongjun Tang, and Rui Zhong. 2024. “The Effects of Mandatory ESG Disclosure Around the World.” *Journal of Accounting Research*. <https://doi.org/10.1111/1475-679x.12548>.
- Lang, Mark, and Lorien Stice-Lawrence. 2015. “Textual Analysis and International Financial Reporting: Large Sample Evidence.” *Journal of Accounting and Economics* 60 (2-3): 110–35. <https://doi.org/10.1016/j.jacceco.2015.09.002>.
- Lawrence, Alastair. 2013. “Individual Investors and Financial Disclosure.” *Journal of Accounting and Economics* 56 (1): 130–47. <https://doi.org/10.1016/j.jacceco.2013.05.001>.
- Leuz, Christian, and Peter D. Wysocki. 2016. “The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research.” *Journal of Accounting Research* 54 (2): 525–622. <https://doi.org/10.1111/1475-679x.12115>.
- Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *34th Conference on Neural Information Processing Systems*.
- Li, Feng. 2008. “Annual Report Readability, Current Earnings, and Earnings Persistence.” *Journal of Accounting and Economics* 45 (2-3): 221–47. <https://doi.org/10.1016/j.jacceco.2008.02.003>.
- Loughran, Tim, and Bill McDonald. 2011. “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” *The Journal of Finance* 66 (1): 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>.
- Macnamara, Brooke N., Ibrahim Berber, M. Cenk Çavuşoğlu, Elizabeth A. Krupinski, Naren Nallapareddy, Noelle E. Nelson, Philip J. Smith, Amy L. Wilson-Delfosse, and Soumya Ray. 2024. “Does Using Artificial Intelligence Assistance Accelerate Skill Decay and Hinder Skill Development Without Performers’ Awareness?” *Cognitive Research: Principles and Implications* 9 (1). <https://doi.org/10.1186/s41235-024-00572-8>.
- Miller, Brian P. 2010. “The Effects of Reporting Complexity on Small and Large Investor Trading.” *The Accounting Review* 85 (6): 2107–43. <https://doi.org/10.2308/accr.00000001>.
- Milne, Markus J., and Rob Gray. 2012. “W(h)ither Ecology? The Triple Bottom Line, the Global Reporting Initiative, and Corporate Sustainability Reporting.” *Journal of Business Ethics* 118 (1):

- 13–29. <https://doi.org/10.1007/s10551-012-1543-8>.
- Noy, Shakked, and Whitney Zhang. 2023. “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.” *Science* 381 (6654): 187–92. <https://doi.org/10.1126/science.adh2586>.
- PwC. 2023. “PwC’s Global Investor Survey 2023.” PwC. <https://www.pwc.com/gx/en/issues/c-suite-insights/global-investor-survey/global-investor-survey-2023.html>.
- Rennekamp, Kristina. 2012. “Processing Fluency and Investors’ Reactions to Disclosure Readability.” *Journal of Accounting Research* 50 (5): 1319–54. <https://doi.org/10.1111/j.1475-679x.2012.00460.x>.
- Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. “When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis.” *Nature Human Behaviour* 8: 2293–303. <https://doi.org/10.1038/s41562-024-02024-1>.
- Wagenhofer, Alfred. 2024. “Sustainability Reporting: A Financial Reporting Perspective.” *Accounting in Europe* 21 (1): 1–13. <https://doi.org/10.1080/17449480.2023.2218398>.
- Weil, David, Mary Graham, and Archon Fung. 2013. “Targeting Transparency.” *Science* 340 (6139): 1410–11. <https://doi.org/10.1126/science.1233480>.
- Yang, Cathy (Liu), Kevin Bauer, Xitong Li, and Oliver Hinz. 2026. “My Advisor, Her AI, and Me: Evidence from a Field Experiment on Human–AI Collaboration and Investment Decisions.” *Management Science* 72 (1): 242–64. <https://doi.org/10.1287/mnsc.2022.03918>.
- Yip, Rita W. Y., and Danqing Young. 2012. “Does Mandatory IFRS Adoption Improve Information Comparability?” *The Accounting Review* 87 (5): 1767–89. <https://doi.org/10.2308/accr-50192>.

Appendix A. Variable definitions

Variable	Definition
Panel A: Field experiment (Section 3)	
<i>SCORE</i>	Composite performance in a round, $z(\text{accuracy}) + z(\text{speed})$, win-sorized and shifted to be non-negative; defined relative to the estimation sample.
<i>CORRECT</i>	Number of correct answers in a round (0–3), summing the three single-choice emissions questions (absolute emissions, emissions intensity, and most ambitious climate target).
<i>TIME (experiment)</i>	Minutes a participant spends analyzing the three reports in a round.
<i>CSRD (experiment)</i>	Treatment indicator equal to one when the participant works with CSRD standardized reports rather than pre-CSRD reports.
<i>AI</i>	Treatment indicator equal to one when the participant has access to the custom retrieval-augmented (RAG)-based AI chatbot over the reports.
<i>ROUND</i>	Order of the industry round completed by the participant (1st–3rd).
<i>EXPERIENCE</i>	First principal-component factor of the pre-experiment questionnaire on prior exposure and familiarity to reporting, based on 7-point Likert scale questions for 'How often are you investing in the stock market?', 'How often are you considering primary sources (e.g., annual reports) for your investment decisions?', 'How often do you rely on investment recommendations from friends and colleagues?', and 'How often do you use generative AI tools?'
<i>FEMALE</i>	Indicator for female participant, as elicited in the background questionnaire.
<i>AGE</i>	Ordered category codes for age bracket (18-24; 25-34; 35-44; 45-54; 55-65; 65+), as elicited in the background questionnaire.
<i>EDUCATION</i>	Ordered category code for highest degree (No formal education; A-levels; High school; Vocational training; Undergraduate; Post-graduate Master's degree; Doctorate), as elicited in the background questionnaire.
<i>USED AI</i>	Indicator for issuing at least one chatbot query in a round.
<i>QUERIES</i>	Number of chatbot queries per round.

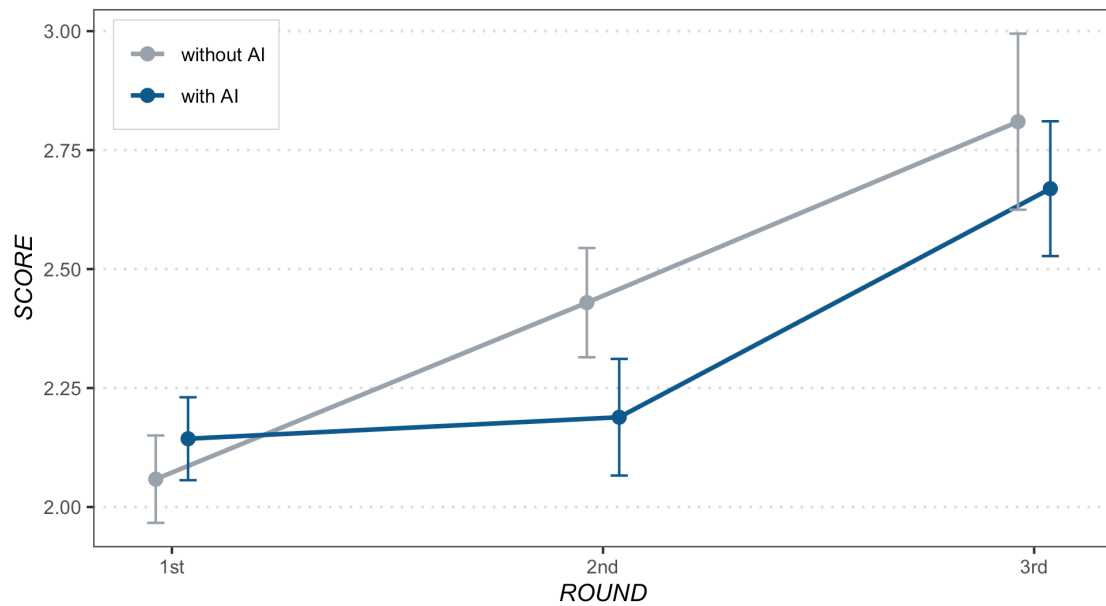
Notes. Variables used in the field experiment (Section 3); measured at the participant-by-industry round.

Variable	Definition
Panel B: Field evidence (Section 4)	
<i>TIME (field)</i>	Minutes a collector takes to record all data points for one firm-year-topic session; sessions exceeding 120 minutes are dropped as breaks.
<i>CSRD (field)</i>	Indicator equal to one for the first mandatory CSRD reports, filed for fiscal year 2024.
<i>DATA POINTS DISCLOSED</i>	Number of data points actually disclosed for the firm-year.
<i>SESSION</i>	Indicator equal to one if a specific data point is disclosed.
<i>EXPERIENCE</i>	Collector's internal experience: an incrementing counter of completed firm-years.
<i>DISPERSION</i>	First principal-component factor of the pre-collection survey on prior experience.
<i>DIFFICULT</i>	Herfindahl–Hirschman index of the page numbers across which a firm-year's data points are located.
<i>TONE</i>	Gunning–Fog readability index of the source pages.
<i>COMPLEX</i>	Loughran–McDonald (2011) net tone (positive minus negative word share) of the source pages.
<i>UNCERTAIN</i>	Loughran–McDonald (2011) complexity-word share of the source pages.
<i>TABLES</i>	Loughran–McDonald (2011) uncertainty-word share of the source pages.
<i>IMAGES</i>	Counts of tables on the source pages for which data points were found.
	Counts of images on the source pages for which data points were found.

Notes. Variables used in the field evidence (Section 4); measured at the collector-firm-year session.

Online Appendix

Figure OA-1: SCORE across successive rounds, by AI access



Notes. Points are the mean of *SCORE* in each round (1st–3rd), shown separately for participants without and with access to AI; whiskers span ± 1 standard error. All variables are defined in Appendix A.

Table OA-1: AI access versus use

Dependent variable:	<i>SCORE</i>	<i>SCORE</i>	<i>SCORE</i>
	(1)	(2)	(3)
<i>AI</i>	-0.048 (0.100)	-0.363*** (0.099)	-0.067 (0.137)
<i>CSR</i> <i>D</i>	0.323*** (0.102)	0.316*** (0.101)	0.322*** (0.102)
<i>ROUND</i>	0.310*** (0.061)	0.311*** (0.060)	0.310*** (0.061)
N	760	760	760
Adj. R^2	0.051	0.070	0.057
Industry FE	Yes	Yes	Yes

Notes. *SCORE* is regressed on *AI*, *CSR**D*, and *ROUND*, with industry fixed effects; standard errors clustered by participant. Column 1 estimates the effect of AI access by intention-to-treat (random assignment); column 2 is OLS on realized AI use; column 3 is 2SLS instrumenting realized use with random assignment. All variables are defined in Appendix A.

Table OA-2: Query and answer quality

Dependent variable:	<i>SCORE</i>	<i>SCORE</i>	<i>SCORE</i>
	(1)	(2)	(3)
<i>Share well-formed queries</i>	-0.308 (0.357)	-0.338 (0.363)	-0.441 (0.374)
<i>Share correct AI answers</i>	-0.003 (0.349)	-0.079 (0.345)	-0.022 (0.342)
<i>CSRD</i>			0.389** (0.169)
<i>ROUND</i>		0.268*** (0.099)	0.284*** (0.101)
N	283	283	283
Adj. R^2	-0.006	0.016	0.036
Industry FE	Yes	Yes	Yes

Notes. OLS of *SCORE* on the share of well-formed AI queries and the share of correct AI answers, adding *ROUND* in column 2 and *CSRD* in column 3; all columns include industry fixed effects; sample restricted to rounds with at least one AI query, with query and answer quality aggregated to the participant-by-industry round from per-exchange LLM-judge ratings; standard errors clustered by participant. All variables are defined in Appendix A.

Table OA-3: Matched data-point evidence

Dependent variable:	<i>DISCLOSED</i>	<i>log TIME/DATA POINTS</i>	<i>log TIME/DATA POINTS</i>
	(1)	(2)	(3)
<i>CSRD</i>	0.172*** (0.016)	-0.341** (0.166)	0.075 (0.140)
<i>DISCLOSED</i>			1.373*** (0.054)
<i>CSRD</i> × <i>DISCLOSED</i>			-0.402*** (0.075)
<i>SESSION</i>	0.000 (0.000)	-0.004*** (0.001)	-0.001 (0.002)
N	112,895	24,144	112,895
Adj. R^2	0.488	0.353	0.373
Cell FE	Yes	Yes	Yes

Notes. Estimates on the matched emission-metric data points, i.e., firm-by-data-point pairs that the same collectors recorded both before the CSRD and under the first mandatory CSRD reports (fiscal year 2024). The unit of observation is the collector's first write of a data point. Column 1 is a linear probability model for whether the item is *DISCLOSED*; columns 2 and 3 regress the natural logarithm of the per-item acquisition time (in seconds, capped at 15 minutes), column 2 on the disclosed items only and column 3 on all matched items with the *CSRD* × *DISCLOSED* interaction. Every column controls for *SESSION* and absorbs a firm-by-data-point cell fixed effect (Cell FE); standard errors are clustered by collector. All variables are defined in Appendix A.