DIR DIGITAL INVESTOR RELATIONS

GenAI as a Reader:
How ChatGPT & Co. use Annual Reports

Authors

Prof. Monika Kovarova-Simecek

USTP – University of Applied Sciences St. Pölten

Prof. Dr. Henning Zülch / Leon Kirschbaum / Konstantin Klammer

HHL Leipzig Graduate School of Management

Dr. Eloy Barrantes / Alexandra Horváthová / Christina Schilling

nexxar

University of Applied Sciences St. Pölten

HHL LEIPZIG GRADUATE SCHOOL OF MANAGEMENT

nexxar

1. Executive Summary

AI tools have become a central gatekeeper between companies and stakeholders. In particular, because an increasing number of investors are using large language models (LLMs) such as ChatGPT to research companies and challenge investment ideas. Research from the University of Washington highlights this shift: nearly half of the 2,175 retail investors surveyed already use generative AI tools to analyze financial and market information.[1]

However, this shift in information usage leads to different risks for listed companies: LLMs such as ChatGPT can hallucinate or generate plausible-sounding but incorrect and unverified answers to stakeholder questions.[2] This is particularly risky when information is derived from external third-party sources, given that sources like Reddit and Wikipedia currently dominate as data sources for LLMs such as ChatGPT or Perplexity.[3] Moreover, especially user-generated content often combines information about a company's development with opinions and framing.

In the age of AI, listed companies have a particular interest in ensuring that their own corporate content is represented in the responses generated by large language models. In this regard, corporate reports – especially Annual Reports – are the most comprehensive and reliable source of information on a company's financial and non-financial development. They undergo internal verification by the company's management and are externally approved by certified auditors. This clearly distinguishes annual reports from the vast majority of unverified information about a company, which users can find online. In theory, the character of "verified information", makes annual reports a perfect source for LLMs: Given that most of the global internet content lacks formal verification, the trustworthiness of corporate reporting documents positions them as potentially valuable inputs for tools such as ChatGPT.

[1] More sophisticated investors are leading the way, applying these tools to more complex tasks – and a strong majority (76%) expect GenAI to become a standard part of the investment process in the future. See Blankespoor/Croom/Grant (2024): Generative AI and Investor Processing of Financial Information.

[2] See Jaźwińska/Chandrasekar: AI Search Has a Citation Problem (2025). The study tested eight generative AI search tools (including ChatGPT, Perplexity, Copilot, Gemini) on news-related queries and found that AI search engines failed to retrieve correct news citations in over 60 % of cases.

[3] See Semrush (2025): How Google's AI Mode Compares to Traditional Search and Other LLMs [AI Mode Study]

Despite this theoretical value, the extent to which corporate reports are accessible for and listed by LLMs remains unclear. Specifically, questions persist about the visibility of different communication formats – such as digital HTML versus static PDF documents. To address these questions, USTP – University of Applied Sciences St. Pölten, HHL Leipzig Graduate School of Management, and nexxar have initiated a joint research effort composed of two related studies.

- **Visibility of Annual Reports in Large Language Models**: The first study was a large-scale ChatGPT experiment examining how annual reports are used as sources in AI-generated answers and whether the format of disclosure – PDF versus structured online reports – affects their visibility. We systematically tested citation patterns to assess how frequently and in what way corporate reporting content is referenced in ChatGPT.
- **AI-usage of Digital Reports**: The second study focused on log file analysis of digital annual reports. Here, we investigated how often Bots and LLMs actually access reports of different stock-listed companies, which sections are visited, and which pages are most relevant from an AI perspective.

Content

2. Visibility of Annual Reports in LLMs

An Empirical Study on Source Usage Patterns of ChatGPT

Prof. Monika Kovarova-Simecek / Dr. Eloy Barrantes / Alexandra Horváthová

2.1 Abstract

Large Language Models (LLMs) such as ChatGPT are increasingly used as intermediaries for accessing corporate financial information. This study investigates to what extent verified corporate reporting information is reflected in AI-generated answers and which source types are preferentially used. A research team[4] conducted more than 2,500 structured prompts related to corporate reporting topics for 20 publicly listed European companies using ChatGPT (GPT-4o and GPT-5). More than 24,000 cited sources were extracted, coded, and analyzed. The results show that ChatGPT predominantly relies on original corporate disclosures, especially annual reports and investor relations websites. However, the format of disclosure plays a decisive role: companies providing structured online (HTML) annual reports were referenced significantly more often than their peers publishing primarily PDF reports. The findings indicate a strong preference of LLMs for semi-structured, machine-readable reporting formats, with implications for corporate disclosure strategies in the age of AI-mediated information retrieval.

2.2 Research Objectives and Research Questions

The primary objective of this study is to analyze which sources Large Language Models like ChatGPT use when answering questions about corporate reporting topics. The study is guided by the following research questions:

RQ1: Which categories of sources are cited by ChatGPT when answering reporting-related prompts about publicly listed companies?

RQ2: Does the format of the annual report (HTML-based online report vs. PDF-based report) influence the visibility of corporate information in AI-generated answers?

[4] Many thanks to the entire research team from USTP – University of Applied Sciences St. Pölten: Adela Danciu, Andreas Hohenauer, Ciara Steurer, Cornelia Isabell Plott, Cornelia Renner, Elena Gehmayr, Evgeniya Niberg, Hannah Hössinger, Jana Beisser, Johanna Neumann, Katarzyna Leduc, Katharina Woloch, Lara Koller, Luisa Storfa, Max Schlatterbeck, Michael Hammerer, Nicolas Kubrak, Rosanna Pospisil, Sandra Kortschak, Simon Kranawetter, Sophie Aimée Horcher, Tanja Maisenberger, Yulia Vandzhura.

2.3 Method

The sample consists of 20 publicly listed companies, across nine European countries (including Germany, Austria, Switzerland, the United Kingdom, and others), and eight industries (such as pharmaceuticals, automotive, telecommunications, chemicals, and aviation). The companies were divided into two equally sized groups: Online Report Group (n = 10): Companies publishing a fully fledged, structured HTML annual report. PDF Report Peer Group (n = 10): Companies publishing primarily PDF-based annual reports. The two groups were constructed as peer groups to ensure comparability.

For this study, more than 2,500 prompts related to corporate reporting topics were submitted to ChatGPT for a sample of 20 publicly listed companies. The sample was divided into two equally sized peer groups to ensure comparability. Group A consisted of companies publishing fully structured HTML-based annual reports (= Online Report Group), while the companies in Group B primarily published their report as a PDF version (=PDF Report Group).

The experiment was conducted using ChatGPT, specifically the GPT-4o and GPT-5 models. A total of 23 academic testers participated in the study. Each tester coded prompts for six companies from the sample, ensuring overlaps, a high level of inter-coder reliability, and cross-validation.

The prompts covered a broad range of corporate reporting topics and were grouped into the following categories:

- Annual Report availability and access
- Financial performance and metrics
- Board of Management and remuneration
- Management Report (strategy, risks, innovation)
- Sustainability and ESG disclosures (ESRS, E1, S1, taxonomy, assurance)

Each prompt was fully standardized and customized only by replacing the company name (e.g. “Does %company% publish an annual report?”, “What are the E1 targets of %company% for 2024?”, “Please provide me with the latest balance sheet of %company% as XLS-file.”).

The study distinguishes between internal sources (annual reports, corporate websites, and presentations) and external sources (news media, financial data aggregators, user-generated content, and social media). These categories were further refined, resulting in a total of 30 distinct source types that were identified and analysed. The study focuses exclusively on explicitly cited sources shown within ChatGPT answers. “Background sources” which were not visible in the direct answer, were not included.

After each coding session, the full chat history was archived for traceability. Sources were classified based on their domain and URL structure into ten predefined categories (see table):

Sources	Description
[SOUR1] = Annual report (online) – internal	All links to the digital annual report (regardless for which financial year). Only sources, which are part of the corporate website URL (e.g., www.companyname.com/XY or report.companyname.com).
[SOUR2] = Annual report (PDF) – internal	All links to the PDF annual report (regardless for which financial year). Only sources, which are part of the corporate website (e.g., www.companyname.com/XY). PDF downloads from other websites were coded as external sources.
[SOUR3] = Investor Relations website – internal	All links (except SOUR1 & SOUR2) which are part of the Investor Relations section of the corporate website. These links can normally be identified with the URL and need to be part of the corporate website (e.g., www.company.com/investors or www.company.com/investor-relations or www.company.com/ir/)
[SOUR4] = Investor presentations and resources (PDF) – internal	All links (except SOUR1-SOUR3) to Investor presentations and resources (PDFs) – normally within the Investor Relations websites
[SOUR5] = Corporate website - internal	All links to the corporate website (company.com), which are not part of the IR section (see [SOUR4]).
[SOUR6] = News media – external	Articles from news outlets (e.g., Financial Times, Handelsblatt): 61 = Handelsblatt 62 = Wall Street Journal 63 = Financial Times 64 = The Times 65 = The Guardian 66 = Other news media
[SOUR7] = Financial data aggregators – external	Financial data aggregators like Yahoo Finance, Morningstar, MarketScreener, etc.: 71 = MarketScreener 72 = Bloomberg 73 = Reuters 74 = Morningstar 75 = Yahoo finance 76 = Annualreports.com 77 = Finanzen.net/Finanzen.at 78 = Other financial data aggregators
[SOUR8] = User Generated Content – external	All sources to User generated content websites – e.g., Wikipedia, Reddit, Investopedia, Slideshare etc.: 81 = Reddit 82 = Wikipedia 83 = Other user generated content platforms
[SOUR9] = Social Media – external	Social Media sites: 91 = LinkedIn 92 = Facebook 93 = YouTube 94 = Instagram 95 = Other social media platforms
[SOUR10] = Other – external	All other external sources

In total, more than 24,000 cited sources were extracted, coded, and analyzed within this study.

2.4 Key Results

Annual Report Visibility in ChatGPT

The results clearly show that corporate reports are the single most important and trusted source category for ChatGPT when it comes to questions on the financial- and non-financial development of listed companies. Out of 24,662 analysed citations, 58.5% were directly linked to the respective company's annual report. This finding confirms the central role of formal corporate reporting even in AI-mediated information retrieval contexts.

Impact of Reporting Format (HTML vs. PDF)

However, our results show a **significant impact of the reporting format** on the AI-visibility of the report: Companies with HTML-based online annual reports (Group A) were cited **three times more frequently (3.05x)** with respect to annual report content than companies relying primarily on PDFs (Group B). In absolute terms, digital (HTML) reports generated significantly more internal citations than PDF-only setups.

Sources	Group A (Online Report)	Group B (PDF Report)
Online Report	9,557	1,172
PDF Report	1,298	2,385
IR website (incl. downloads)	862	2,921
Corporate website	1,116	1,778
Media articles (e.g. Financial Times)	208	653
Financial data aggregators (e.g. Market-Screener)	335	1,171
User generated content (e.g. Wikipedia)	7	16
Social Media (e.g. LinkedIn)	1	13
Other external sources	408	761

N = 24,662 citations in ChatGPT

In contrast, PDF-only reporting setups lead to a substantially higher reliance on external sources. The analysis shows that answers related to PDF reporters contained approximately 2.7 times more external citations than those related to digital reporters. Financial data aggregators such as MarketScreener, Bloomberg, or Reuters emerged as the most significant competitors to corporate-owned sources in these cases.

This shift towards external intermediaries implies a loss of narrative and factual control, as AI systems increasingly depend on third-party interpretations rather than primary disclosures.

Accuracy of Answers

Beyond visibility and source attribution, the study also assesses the factual correctness of ChatGPT responses. A dedicated subsample analysis (n = 200 responses) reveals a clear accuracy gap between digital and PDF reporters. Responses related to companies with HTML reports (Group A) were correct in 71% of cases, compared to 54% for companies relying on PDF-only reports (Group B). The higher accuracy is attributed to better accessibility, clearer structure, and reduced dependence on secondary sources.

2.5 Discussion

The findings demonstrate a clear preference of ChatGPT for structured, machine-readable corporate disclosures. The reporting format (HTML vs. PDF) does not merely influence visibility, but also materially affects the quality and reliability of AI-generated information. From a corporate perspective, this has direct implications for investor communication, reputation management, and the risk of misinformation.

HTML-based annual reports provide semi-structured data that LLMs seem to process, interpret, and extract more effectively than static PDF documents. While the PDF annual reports of all companies in the sample were also cited by ChatGPT, digital reports appeared in the responses nearly three times as often. Moreover, responses concerning companies with HTML-based reports relied on significantly fewer external sources.

The increased reliance on external sources in the PDF group suggests a potential risk for companies: when original disclosures are less accessible to AI systems, third-party interpretations gain relative importance. This may lead to reduced control over how corporate information is represented in AI-mediated environments.

3. LLMs as Users of Digital Annual Reports

Prof. Dr. Henning Zülch / Leon Kirschbaum / Konstantin Klammer / Christina Schilling

3.1 Abstract

The growing use of artificial intelligence to access and analyze corporate digital information is fundamentally changing how digital annual reports are used. This study examines usage patterns of digital annual reports based on server-side access log data from five DAX-listed companies. The empirical dataset comprises more than 4.8 million automated requests, of which 759,226 bot requests were included in the content-level analysis after data cleaning. The results show that automated systems account for a substantial share of total access activity, with more than 175 identifiable bots interacting with digital reports. Usage is highly concentrated: the five most active bots generate 48.8% of all automated requests, and ChatGPT alone accounts for 30.3% of total bot traffic. In terms of content, access activity focuses primarily on core report sections, particularly operating activities, strategic direction, followed by information on financial statements and key financial metrics, and sustainability reporting. Access patterns also reveal a clear preference for the most recent reporting periods. Overall, the study provides new empirical insights into the role of artificial intelligence as a dominant user of digital annual reports and discusses implications for the design and access management of corporate reporting in the age of AI.

3.2 Research Objectives and Research Questions

The objective of this study is to provide empirical evidence on how bots access digital annual reports. In particular, the dataset allows us to see which LLMs are the most dominant and which categories in the annual report are the most analyzed by LLMs.

RQ1: How is access to digital annual reports distributed across different bots and which actors dominate overall access activity?

RQ2: Which report contents and reporting periods of digital annual reports are most frequently accessed?

3.3 Method

This study is based on the analysis of server-side access data from digital annual reports. The objective of the empirical analysis is to systematically capture the volume, structure, and content focus of the use of digital corporate reporting. The empirical dataset consists of server log files from five companies listed in the DAX index, which record all requests to the companies' corporate websites and the subpages of their digital annual reports. The log files were captured for a total of eight weeks from 29 August 2025 to 25 October 2025. In addition to the accessed URLs, the log files include time stamps, allowing inferences about the reporting periods from which the requested information originates.

For the first research question, bots were identified and aggregated by bot name. In a first step, the number of distinct bots accessing the digital annual reports during the observation period was determined, and bot activity was described using descriptive statistics. In addition, the concentration of bot activity was analyzed by calculating the percentage share of each bot in the total volume of requests. This approach allows for insights into the distribution of automated usage and the identification of particularly dominant actors.

For the second research question, the raw data was subjected to a multi-stage data-cleaning process. First, identical bot requests were aggregated at the URL level. To ensure the statistical robustness of the analysis, only report subpages that recorded at least 100 accesses during the observation period were included. The resulting cleaned dataset comprises a total of 759,226 bot requests, with access volumes varying across the analyzed companies.

Each request was assigned to a thematic reporting category, including financial reporting, operational business activities, sustainability reporting, and shareholder-related information. This classification allows for the examination of which report categories were accessed most frequently during the observation period. The hierarchical structure of the websites further enables analyses at different levels of aggregation, ranging from individual subpages to broader thematic clusters.

3.4 Key Results

Distribution and Dominance of Actors

Automated access to digital annual reports is highly concentrated. Across the observation period, a total of 4,838,833 automated requests were recorded, with more than 175 identifiable bots interacting with digital reports. Bot activity follows a strongly skewed distribution: the five most active bots account for 48.8% of total automated requests, while the majority of bots exhibit only marginal activity. ChatGPT alone generates 30.3% of total bot traffic, followed by Bingbot (6.6%), Amazonbot (5.8%), Baiduspider (4.6%), and Googlebot (4.6%). At the company level, two firms account for 67.7% of all automated requests, indicating an evident concentration of bot activity across digital reports. Overall, the findings point to a small number of dominant actors acting as primary intermediaries in the use of digital annual reports.

Content Focus and Usage Patterns

The analysis of the cleaned dataset (N = 759,226 bot requests) reveals a clear concentration of access on economically relevant core sections of digital annual reports. Requests related to operational business performance account for 39.3% of all analyzed accesses, followed by financial reporting and financial statements with 20.6%. Additional access activity is directed toward strategic and sustainability-related information. Access patterns further indicate a strong preference for current reporting periods. At the firm level, automated usage is unevenly distributed across the five DAX-listed companies, suggesting that content structure, website characteristics and overall public interest, rather than firm size, shape access intensity.

3.5 Discussion

This study demonstrates that digital annual reports are no longer consumed exclusively by human users but are increasingly accessed and processed by large language models. This development effectively expands the readership of corporate reporting to include non-human intermediaries whose access patterns, selection mechanisms, and modes of information reuse differ fundamentally from traditional, manual information consumption. The empirical findings show that automated access constitutes a substantial share of total usage and is highly concentrated: although more than 175 distinct bots interact with digital annual reports, a small number of actors account for the majority of automated requests. In particular, the dominance of ChatGPT highlights the growing importance of generative AI systems as key access points and distribution channels for corporate information. The content-related access patterns further indicate that automated systems focus primarily on report sections with high decision relevance. Requests concentrate on information related to operational performance and financial reporting, followed by strategic and sustainability-related disclosures, and exhibit a strong preference for the most recent reporting periods.

These findings have several implications for both research and practice. First, the strong concentration of access implies that a small number of platform providers increasingly assume a gatekeeper role in shaping the visibility and usability of corporate reporting content. This raises governance-related questions concerning

information control and source authority. Firms are therefore challenged to structure their digital reporting in ways that remain accessible and meaningful for human users while also being clearly interpretable, consistently referenceable, and contextually robust for AI systems. Second, the high share of non-identifiable bots points to transparency deficits in the current digital reporting environment, complicating the monitoring and management of data flows. Third, the study opens several avenues for future research in corporate reporting and capital markets. In particular, the growing role of AI systems as upstream information filters raises important questions about how machine-mediated information processing affects information asymmetries, analyst behavior, media coverage, and ultimately capital market reactions.

Overall, the findings underscore that the transformation of corporate reporting is not solely driven by the introduction of new content areas, such as sustainability reporting, but increasingly by the technical and semantic compatibility of reporting formats with AI-driven information ecosystems. This shift calls for a broader understanding of corporate reporting and a redefinition of who, or what, constitutes the “reader” of an annual report.

Contact

Prof. Dr. Henning Zülch

HHL Leipzig Graduate School of Management

henning.zuelch@hhl.de

www.hhl.de

Prof. Monika Kovarova-Simecek

USTP – University of Applied Sciences St. Pölten

Monika.Kovarova-Simecek@fhstp.ac.at

www.ustp.at

Dr. Eloy Barrantes

nexxar

eloy.barrantes@nexxar.com

www.nexxar.com