The First Annual Reports of the LLM Era

An exploratory analysis of language change in 150 SEC 10-K filings, 2019-2024

Richard Foster-Fletcher

MKAI

Correspondence: research@mkai.org

Revised August 2026

Abstract

This study examines purposively selected narrative passages from 150 SEC Form 10-K filings for a fixed sector-stratified sample of 50 companies. Each company contributes three observations labelled FY2019, FY2022 and FY2024. A single author scored six directional dimensions of language change on an ordinal scale from zero to two. Five dimensions cover changes in claims, hedging, framing, named specificity and sentence expansion. The sixth covers newly introduced or reorganised sections and remains outside the principal editorial comparison.

The score allocation gives 106 points across the five editorial dimensions during 2019-2022 and 88 points during 2022-2024. Dividing those totals by nominal intervals of three and two years produces annual aggregate rates of 35.3 and 44.0 points, an increase of 24.5 per cent. The mean company difference of 0.173 points per company-year equals the aggregate rate difference divided by 50 and supplies no separate estimate. The sample standard deviation is 0.994, the median company difference is 0.000 and a company-resampling bootstrap interval extends from approximately -0.10 to 0.45. A sign test gives p = 0.883. Under the nominal interval convention, later rates are higher for 24 companies, earlier rates are higher for 22 and four are equal. Elapsed fiscal year ends produce counts of 24, 25 and one, while filing dates produce 25, 24 and one. Neither interval produces a majority under any convention.

The result depends upon interpretive scoring, purposive passage selection and the allocation of scores between the two intervals. Moving 10.4 points from the later interval to the earlier interval would equalise their annual aggregate rates, so eleven whole points would reverse their ordering. Alternative codings of the ordinal scale move the aggregate increase from 17.3 to 35.6 per cent, while leaving out one company moves it from 18.9 to 36.1 per cent. The study therefore describes the scored sample and rubric. It does not establish a population trend, identify generative AI use or support causal attribution.

Research purpose

The FY2024 observations were the first in the study collected after enterprise versions of large language model tools became widely available. The study examines whether selected filing passages changed between FY2019 and FY2024, and whether the directional scores accumulated at different annual rates on either side of the FY2022 observation.

The design observes three filing years. It contains no company-level evidence showing whether a large language model contributed to drafting, review or revision. Corporate disclosure also had documented long-term trends before generative AI became available. Dyer, Lang and Stice-Lawrence found increasing 10-K length, boilerplate, stickiness and redundancy between 1996 and 2013, alongside decreasing specificity, readability and hard information. They attributed much of the increase in length to new SEC and FASB requirements.[1] Cao, Jiang, Yang and Zhang subsequently found that growing machine readership affected

how firms prepared disclosure, using machine downloads, investor exposure and earlier natural-language-processing developments to support attribution.[2]

Those long-term findings limit the interpretation of a three-year comparison. A post-2022 observation may continue an older disclosure trend, reflect a common event or mark a new change. The present dataset cannot separate those possibilities.

Sample and passage selection

The study uses a fixed sector-stratified sample of 50 companies across financial services, technology, healthcare, energy and industrials. The companies were selected using market capitalisation from a ranking whose source, date and inclusion rules are unavailable. Each company contributes three observations labelled FY2019, FY2022 and FY2024. The company list was fixed before scoring began.

The researcher selected passages purposively. Item 1 extracts focused on opening business descriptions and identity claims. Item 1A extracts included opening framing and selected risk factors that supported comparison across the three years, together with some newly emerging topics considered diagnostically useful. Item 7 extracts focused on opening narrative discussion of performance, strategy or operating conditions. Within each chosen passage, extraction followed a fixed stopping rule. The available extraction material does not identify which risk factors were chosen for each company, so those selections cannot be reconstructed. The method targeted approximately 15 to 20 pages of narrative material from each filing.

This approach introduced researcher judgement before scoring. Another selection from the same filings could produce different scores. Passage choice and boundary placement were separate stages: selection was purposive, while extraction within each chosen passage followed the fixed stopping rule.

Some primary Form 10-K documents incorporate substantive annual-report text from Exhibit 13, while other filings distribute a legal item across cross-referenced pages. The source manifest identifies the primary and incorporated documents used in the corpus, although the scoring evidence does not link every analysed passage to a precise file boundary. Eight incorporated annual reports supply fourteen verified Item 1A or Item 7 sections, together with one cross-referenced Item 1C subsection.

Two company-items change source document type across the study years. Pfizer Item 7 contains 43,732 verified words from Exhibit 13 in FY2019, followed by 16,095 and 16,413 words from the primary Form 10-K in FY2022 and FY2024. RTX Item 7 contains 18,313 words from an incorporated annual report in FY2019, followed by 20,787 and 21,445 words from the primary filings. A source change and a language change cannot be separated for either company-item. These source issues leave the score arithmetic unchanged while limiting textual comparison and corpus completeness.

Scoring and interval allocation

Six dimensions were scored from zero to two. A zero indicates that the selected language remained stable or moved away from the direction defined by the rubric. A two indicates marked movement in that direction. The scale measures directional change from the selected FY2019 passage. It does not independently measure disclosure quality or information value.

The five editorial dimensions measure the following movements:

1. D1 measures softening or removal of distinctive and testable claims.
2. D2 measures expansion of qualifying or uncertain language.
3. D3 measures separation of general framing from concrete examples.

4. D4 measures replacement of named specifics with more general wording.
5. D5 measures expansion of comparable passages without a proportionate increase in specific content, as judged by the author.

D5 includes an interpretive judgement about proportionate content. The study did not define or calculate independent information gain, entropy or decision value. D5 represents a scored assessment of sentence expansion and specificity within selected comparisons.

D6 concerns new or reorganised sections. SEC cybersecurity rules required annual Form 10-K disclosure concerning cybersecurity risk management, strategy and governance for fiscal years ending on or after 15 December 2023.[3] That change concentrates new Item 1C text in the later period. D6 is reported separately as a sensitivity result.

The company and dimension scores were assigned before allocation between Period A, covering 2019-2022, and Period B, covering 2022-2024. A score could fall wholly within one interval, split between both intervals or require an interpretive allocation where all three passages differed. The allocation protocol records sixteen source-based cell checks and identifies McKesson D2 and Morgan Stanley D2 as inferential. The individual rationales and evidence tables for those checks are unavailable, so the affected cells cannot be identified. The company table labels 35 rows as clean, 13 as partial and two as inferential. The interval composition therefore contains additional researcher judgement beyond the company-level score.

The observed interval result

Across D1 to D5, Period A contains 106 points and Period B contains 88 points. Their unequal lengths require a denominator. Dividing by three years and two years gives 35.3 aggregate points per year during Period A and 44.0 during Period B. The later annual aggregate rate is 24.5 per cent higher under this calculation.

The aggregate calculation treats ordinal points as additive units. Recoding a score of two as 1.5 produces an increase of 29.4 per cent, recoding it as three produces 17.3 per cent and treating every positive score as one produces 35.6 per cent. These three scenarios show an 18.3 percentage-point range without changing any scoring decision, although they do not define universal bounds across every possible recoding.

Expressed per company, the aggregate rate difference is 0.173 points per company-year. The mean of the 50 paired company differences is arithmetically identical because every company uses the same three-year and two-year denominators. It provides dispersion information through a sample standard deviation of 0.994, while the median company difference is 0.000. A company-resampling bootstrap interval extends from approximately -0.10 to 0.45. An exploratory paired t-test gives $p = 0.224$, a sign test gives $p = 0.883$ and a Wilcoxon signed-rank test gives $p = 0.150$. The ordinal scale, purposive sample and interval allocation restrict the interpretation of these diagnostics.

Under nominal three-year and two-year denominators, annualised company rates are higher during Period B for 24 companies and during Period A for 22 companies. Goldman Sachs, Humana and SLB each carry three Period A points and two Period B points, so they tie exactly under those nominal denominators. Elapsed fiscal year ends move all three towards Period A by approximately 0.00046 points per company-year, producing counts of 24 later, 25 earlier and one equal. Filing dates produce 25 later, 24 earlier and one equal.

NVIDIA is equal because it has no scored movement in either interval. Across the tested conventions, each period reaches at most 25 companies.

Transferring 10.4 points from Period B to Period A would equalise the aggregate annual rates. Scores use whole points, so eleven transferred points would make the earlier annual rate higher. Taking the largest Period B cells first, six cells carrying twelve points would reverse the ordering. The allocation protocol

contains sixteen source-based cell checks whose individual rationales are unavailable. This calculation tests how many transferred points would change the aggregate result without assessing any individual allocation.

Composition of the scores

Table 1 reports the dimension allocations. Sentence expansion contributes the largest share of Period A scores. Hedge expansion and sentence expansion have equal Period B totals, with framing close behind.

Table 1. Dimension profile by period, D1 to D5

Dimension	Period A points	Period B points	Period A rate	Period B rate	Rate change
D1 Distinctive claims	25	14	8.3	7.0	-16%
D2 Hedge expansion	23	22	7.7	11.0	+43%
D3 Framing and examples	19	20	6.3	10.0	+58%
D4 Named specifics	8	10	2.7	5.0	+88%
D5 Sentence expansion	31	22	10.3	11.0	+6%
Total	**106**	**88**	**35.3**	**44.0**	**+24.5%**

The table describes how the scores were allocated. It cannot establish that the underlying prevalence of each linguistic feature changed by the reported percentages. Passage selection, score thresholds and interval placement all enter those values.

Adding D6 changes the arithmetic substantially. Period A rises to 107 points and Period B rises to 134 points. Their annual rates become 35.7 and 67.0, producing an 87.9 per cent increase. The concentration of mandatory Item 1C disclosures in FY2024 makes that result unsuitable as an editorial estimate.

Sector pattern

The annualised score rates differ across the five sector groups, and each estimate is sensitive to a single company. Table 2 therefore reports the leave-one-company-out range beside every point estimate. Energy crosses zero under this check, while the six-company industrials estimate moves from 100 to 312.5 per cent.

Table 2. Annualised D1-D5 score rates by sector

Sector	Companies	Period A rate	Period B rate	Rate change	Leave-one-out range
Financial services	12	11.7	8.0	-31%	-47.1% to -7.7%
Technology	12	7.0	13.0	+86%	+57.1% to +140.0%
Healthcare	12	9.3	12.5	+34%	+12.5% to +50.0%
Energy	8	5.3	5.0	-6%	-20.0% to +3.8%
Industrials	6	2.0	5.5	+175%	+100.0% to +312.5%

These groups contain between six and twelve companies. Their estimates inherit every limitation of the overall scoring and remain descriptive subgroup observations. The five sector estimates and five dimension estimates are exploratory comparisons without multiplicity adjustment.

Timing and competing explanations

OpenAI released GPT-4 in March 2023 and ChatGPT Enterprise in August 2023.[4] Those dates make LLM-assisted drafting possible during part of the later interval. They do not establish when any sampled company acquired, approved or used a drafting tool.

The exposure window also differs across fiscal calendars. Microsoft, Apple, NVIDIA, Oracle, Salesforce, Cisco, McKesson and Deere use reporting calendars ending outside December. Salesforce closed FY2024 in January 2024 and McKesson closed it in March 2024, leaving shorter periods after the 2023 enterprise releases. Johnson & Johnson presents a separate source-year issue. Its three sampled fiscal year ends are 29 December 2019, 2 January 2022 and 29 December 2024, while the middle filing reports DocumentFiscalYearFocus 2021 under the FY2022 study label. Its elapsed intervals are 2.01 and 2.99 years against nominal denominators of three and two. The aggregate calculation uses the study labels and nominal denominators. The elapsed-year and filing-date company counts use Johnson & Johnson's recorded dates mechanically, while its contribution receives no substantive interpretation because the middle filing reports FY2021. These fiscal calendars preclude a uniform 12-to-15-month exposure assumption across the sample.

Period A contains the onset and continuation of the COVID-19 pandemic. Companies added risk language, operational qualifications, supply-chain discussion and uncertainty during that interval. The study did not separately classify pandemic-related passages. COVID could increase earlier-period scores, alter their composition or change passage comparability. The direction and size of that influence remain unmeasured.

Regulation can affect every scored dimension. New disclosure requirements can increase length, introduce general framing, add qualifying language and reorganise sections. Dyer and colleagues found that a small number of regulatory topics explained much of the earlier long-run increase in 10-K length and other textual changes.[1] Removing D6 therefore addresses one direct structural change without removing regulatory influence from D1 to D5.

Corporate events create further alternatives. Acquisitions, divestitures, restructurings, changes in legal advice, new management, changing risk exposure and revised filing software can all alter text. The present design does not estimate their separate contributions.

Company score distribution

The score matrix has a mean company score of 4.82, a median of 5 and a sample standard deviation of approximately 1.96. Thirty-one companies score five or above. Eight companies score two or below: American Express, Caterpillar, Deere, Intel, Johnson & Johnson, Valero and ExxonMobil score two, while NVIDIA scores one.

Low scores indicate less movement in the directions defined by the rubric. They cannot establish stable or improved disclosure quality because a zero combines stability with movement towards greater specificity. The study also contains no evidence connecting low scores with stronger editorial control.

Interpretation

Applying the scored allocations and nominal three-year and two-year denominators produces a 24.5 per cent difference between aggregate annual score rates. The paired mean restates the same aggregate difference per company, while the median company difference is zero and no timing convention gives either interval a company majority. Leaving out one company moves the aggregate estimate from 18.9 to 36.1 per cent. Leaving out one editorial dimension moves it from 17.2 to 37.0 per cent. These ranges cover only the specified checks and do not define universal bounds.

The source and allocation evidence narrows the interpretation further. Pfizer and RTX change Item 7 source type across the study years, Johnson & Johnson contributes a middle filing from FY2021 under an FY2022 label, and the per-cell allocation rationales are unavailable. The 24.5 per cent figure is a descriptive output of the scoring, source and denominator choices. Its relation to a typical company or a wider disclosure trend is unsupported.

The study provides an exploratory application of a directional scoring framework to selected filing passages. It identifies candidate measures and source questions for a longer measurement design. Its evidence does not support claims that filing language changed because of generative AI, that internal editorial governance explains company variation or that the measured directions represent declining information value.

A longer annual panel, complete and comparable narrative sections, reproducible linguistic measures and independently validated extraction boundaries are required to determine whether the post-2022 observations represent a break within the older evolution of corporate disclosure.

Data availability

The score matrix, company-level interval table and source manifest are available from the author. The source manifest identifies 150 primary Form 10-K documents and eight incorporated annual reports, with EDGAR addresses, local paths, checksums and byte counts. The incorporated documents supply fourteen verified Item 1A or Item 7 sections and one cross-referenced Item 1C subsection. Verified narrative boundaries form part of a separate corpus reconstruction and were not part of the scoring evidence used here.

Use of AI

Large language models assisted extraction, scoring, checking, analysis and drafting. ChatGPT using OpenAI GPT-5.4 assisted initial extraction and preliminary scoring. Claude 4.6 Opus assisted calibration, passage checking, interval allocation and drafting. OpenAI GPT-5.6 generated deterministic verification code that reproduced the score and allocation arithmetic and calculated the additional diagnostics. A later Claude model, whose precise version was not retained, conducted a methodological review of the calculations and claims. Because the Claude model family assisted both interval allocation and methodological review, agreement between those stages carries no independent evidential weight. The numerical claims rest on the available CSV files and reproducible calculations. The author reviewed the evidence, determined the claims and accepts responsibility for the paper.

References

[1] Dyer, T., Lang, M. and Stice-Lawrence, L. (2017). "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation." *Journal of Accounting and Economics*, 64(2-3), 221-245. https://doi.org/10.1016/j.jacceco.2017.07.002

[2] Cao, S., Jiang, W., Yang, B. and Zhang, A. L. (2023). "How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI." *The Review of Financial Studies*, 36(9), 3603-3642. https://doi.org/10.1093/rfs/hhad021

[3] US Securities and Exchange Commission (2023). "Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure." Release Nos. 33-11216 and 34-97989. https://www.sec.gov/rules-regulations/2023/07/s7-09-22

[4] OpenAI (2023). "GPT-4," 14 March 2023; "Introducing ChatGPT Enterprise," 28 August 2023. https://openai.com/index/gpt-4-research/ and https://openai.com/index/introducing-chatgpt-enterprise/

Appendix A. Sample companies

Financial services: JPMorgan Chase, Bank of America, Goldman Sachs, Citigroup, Wells Fargo, Morgan Stanley, American Express, Capital One, US Bancorp, PNC Financial, BlackRock and Charles Schwab.

Technology: Microsoft, Apple, Alphabet, NVIDIA, Tesla, Meta Platforms, Amazon, Oracle, Salesforce, Intel, AMD and Cisco.

Healthcare: UnitedHealth, Eli Lilly, McKesson, Bristol-Myers Squibb, Humana, Merck, AbbVie, CVS Health, Cigna, Elevance Health, Pfizer and Johnson & Johnson.

Energy: Occidental Petroleum, SLB, Marathon Petroleum, Chevron, ConocoPhillips, Phillips 66, ExxonMobil and Valero Energy.

Industrials: GE Aerospace, Honeywell, UPS, RTX Corporation, Caterpillar and Deere.

Appendix B. Scoring rubric

Dimension	Score 0	Score 1	Score 2
D1 Distinctive claims	Claims remained stable or became more specific	Minor softening of testable or comparative claims	Marked removal or replacement of testable identity language
D2 Hedge expansion	No increase in qualifying language	Moderate expansion of conditional wording	Marked or meaning-changing expansion of conditional wording
D3 Framing and examples	Framing remained connected with concrete evidence	Some general framing appeared before examples	A clear pattern of general framing preceded specifics
D4 Named specifics	Named entities, statutes and metrics were retained or increased	Some named specifics were replaced by general wording	Named references were repeatedly replaced by general wording
D5 Sentence expansion	Comparable passages remained stable, became shorter or added proportionate specific content	Passages expanded with some loss of specificity or added qualifying material	Passages expanded markedly without a comparable increase in specific content
D6 New sections	New sections resembled legacy prose, or no new section appeared	New sections appeared somewhat more general than legacy prose	New sections appeared markedly more general or templated