Longitudinal Monitoring and Evaluation of ESG Disclosures from Archival Web Data

Anonymous ACL submission

Abstract

While recent NLP methodologies evaluate corporate sustainability texts, they typically analyze isolated documents without external grounding against quantitative impact estimates. We introduce a pipeline for longitudinal monitoring of unregulated impact disclosures under evolving web structure. Using adaptive filtering and temporal reconstruction over historical web archives, we construct a novel corpus of 24,947 unique documents arranged as 109,759 quarterly firm-state files across 436 private market firms. A Large Language Model (LLM) scores impact categories with high agreement against a domain expert ($r = 0.850$). Contrasting these disclosure scores against external portfolio-weighted impact estimates reveals disclosure-impact divergence ($\rho = 0.118$). Our results demonstrate how longitudinal NLP pipelines support scalable monitoring of corporate impact communication beyond static document analysis.

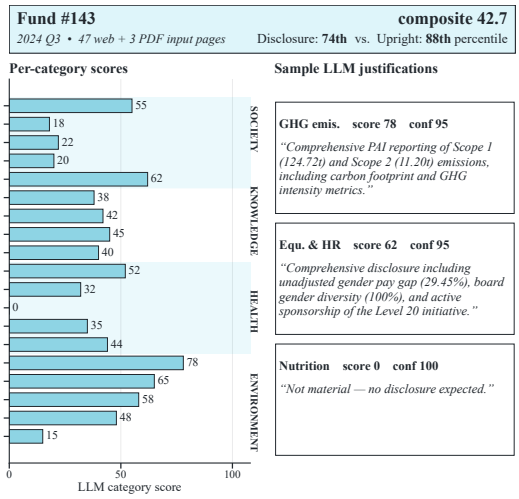

Figure 1: Example output for a randomly-selected fund (index 143 in the dataset). The panel on the left shows the per-category scores, the panel on the right the justifications for three of the presented scores. The bar at the top shows the final composite score, as well as the difference in percentile in this paper's disclosure ranking vs. the Upright Platform.

1 Introduction

Impact investment funds increasingly allocate capital toward measurable societal and environmental solutions, yet private market funds often operate with limited disclosure mandates. Consequently, investors and regulators have limited visibility into how these firms communicate and justify the societal impact of their portfolios. While platforms like the Upright Platform (The Upright Project, 2026, 2020) benchmark corporate value using proprietary metrics (Cheong et al., 2022), their underlying data remains partially inaccessible. In practice, stakeholders rely on unregulated corporate websites and reports, which may not accurately reflect actual investment performance. Crucially, this asymmetry is not static: disclosure practices evolve quarter-over-quarter under regulatory phase-in (SFDR Art. 8/9 (European Union, 2019), CSRD FY2024+ (European Union, 2022)), creating a need not for a single audit but for a continuously refreshed monitoring instrument that tracks how firm communication drifts relative to independent impact estimates.

Furthermore, disclosure practices evolve continuously in response to changing regulation and website redesigns. Analogous to the more common term of greenwashing, this paper describes *impact washing* as claims of investing capital to combat environmental, societal, health, and knowledge-related problems without providing evidence for substance within the underlying portfolios.

We address this issue through a scalable, longitudinal NLP pipeline mapped in Figure 2. Unlike prior ESG NLP work centered on static reports, our framework performs continuous quarter-over-quarter monitoring using archival web data. We contribute: (i) a quarterly-refreshable lon-

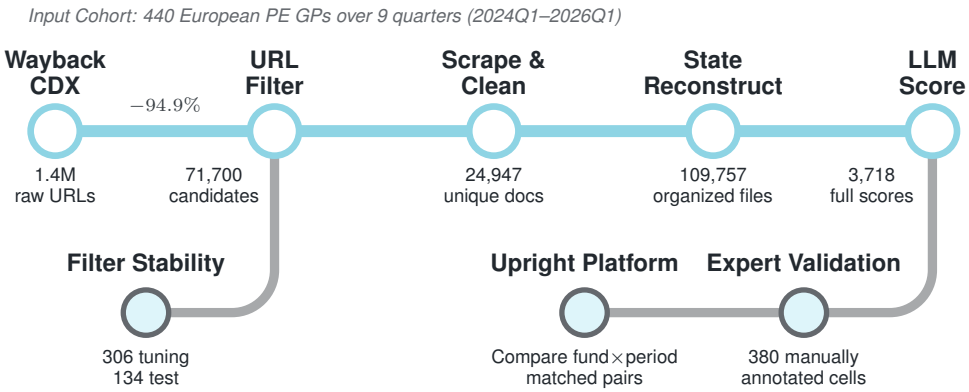

Figure 2: Visual mapping of the data engineering and evaluation pipeline. The primary pipeline (upper track) delineates the sequential transformation of historical web data. Evaluation (lower track) isolate specific pipeline phases, assessing the stability of the URL filter, benchmarking the final LLM output against 380 human annotations and comparing to external Upright-database pairs.

gitudinal pipeline released alongside its first 9-quarter snapshot ($\sim$ 24,947 unique documents, 109,759 quarter-state files, 436 firms), demonstrated to generalize to unseen firms and designed for continuous future-quarter republication; (ii) a schema-constrained LLM evaluation framework with reason-coded scoring, validated against a domain expert at $r = 0.850$; and (iii) an externally grounded panel analysis showing that disclosure scores correlate only weakly with independent portfolio-weighted impact estimates ($\rho = 0.118$). All datasets, prompts, and analysis scripts are released to support replication.[1]

2 Related Work

Sustainability NLP and Web Archives. Existing NLP research on corporate sustainability predominantly evaluates static, standardized annual reports in public equity markets (Webersinke et al., 2021; Chuang et al., 2025). This point-in-time approach translates poorly to private markets, where disclosures are voluntary, fragmented across websites, and continuously evolving. While web archives are established sources for historical NLP corpora (Scialom et al., 2020; Braun and Oswald, 2025) and dictionary-based financial research (Abraham et al., 2024), tracking semantic shifts in ESG reporting requires specialized longitudinal pipelines. Specifically, monitoring these environments involves adapting extraction heuristics to overcome structural web volatility.

[1]A separate repository with full code and data will be provided upon publication.

LLM Evaluation and External Grounding. To evaluate qualitative text, recent work increasingly adopts LLMs as zero-shot judges (Gilardi et al., 2023; Willard and Louf, 2023). However, current ESG NLP frameworks typically measure taxonomy alignment in isolation, lacking external grounding against quantitative performance metrics. While the financial literature notes the prevalence of selective disclosure (Marquis et al., 2016; Gibson Brandon et al., 2022), automated text analysis often remains decoupled from real-world indicators.

3 Methodology & Experiments

Our pipeline transforms public web presences for an input cohort of 436 European private-market GPs (fund managers) into a longitudinal panel of impact-disclosure scores (while text-derived disclosure scores are computed at the GP level, the external impact estimates are evaluated at fund level). The released repository contains the complete filtering vocabulary (Appendix A.1) and scripts required to reproduce all reported experiments.

3.1 Quarterly Website Reconstruction

We query the Internet Archive's CDX index for `text/html` and `application/pdf` snapshots.

(i) Candidate document selection. A heuristic retrieval filter isolates substantive disclosure pages from navigation and boilerplate content. Linked PDFs are extracted and OCR processed.

(ii) Quarterly normalization. We retain one representative snapshot per quarter for each retained document to reduce uneven archival recrawl frequency.

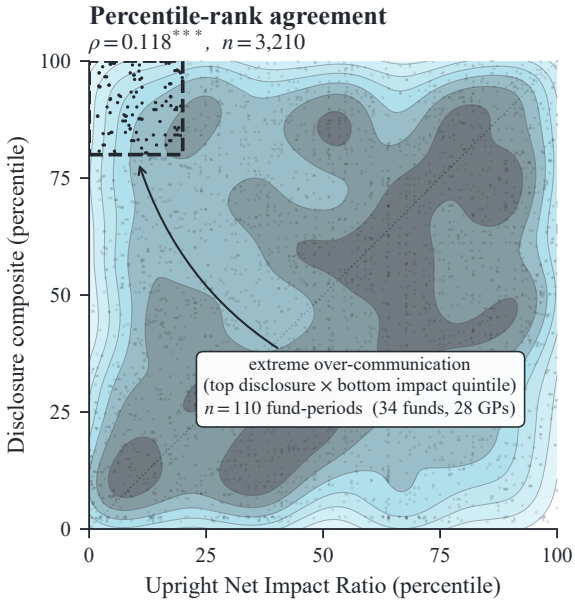

Figure 3: Bivariate density distribution of LLM-derived disclosure percentiles versus actual Upright Net Impact percentiles. The highlighted top-left region isolates a tail of extreme misclassifications (top-quintile disclosure coupled with bottom-quintile actual impact).

(iii) Temporal reconstruction. We reconstruct document lifetimes using a Last Observation Carried Forward (LOCF) assumption. Prior text is retained until explicit deletion (404/410) or prolonged disappearance from the archive.

3.2 Scoring and Evaluation Framework

Each reconstructed firm-quarter corpus is scored by Gemini 3 Flash (Team et al., 2023) using a frozen, expert-authored rubric of 19 categories across four impact pillars (*Society*, *Knowledge*, *Health*, *Environment*; Appendix A.4). Categories range from concrete regulatory anchors (*GHG emissions*, *taxes*) to softer constructs (*meaning & joy*, *relationships*). For each, the model emits a 0–100 score and rationale (Figure 1).

Materiality-conditioned composite. Category scores aggregate into a composite comp $= d \times b$. Depth $d = \frac{1}{n_a} \sum_{c \in \mathcal{A}} s_c$ averages scores across the n_a *addressed* categories $\mathcal{A}$ ($s_c > 0$). Breadth $b = n_a/n_r$ scales this by the number of *materially relevant* categories (n_r). To determine n_r, zero-scored categories require a reason code. Categories tagged *not material* are excluded from n_r, whereas *material but undisclosed* or *insufficient disclosure* categories remain, depressing breadth. This materiality-conditioned denominator distinguishes silent omission from non-applicability. It rewards substantive disclosure on strategically relevant topics without penalizing justifiable omissions, distinguishing our framework from ESG scoring approaches that aggregate uniformly over a fixed taxonomy.

Evaluation Setup. We evaluate the pipeline via three protocols: (i) To compare disclosures against real-world impact, we contrast the composite scores against the Upright Project's *Net Impact Ratio*. Scores are converted to percentile ranks within each cohort. (ii) To validate scoring accuracy, a domain expert re-scored 380 cells. (iii) To test filter stability, we split the GPs into a tuning set ($n_{\text{tune}} = 306$) and a held-out test set ($n_{\text{test}} = 134$).

4 Results

4.1 Disclosure Is Decoupled from Real-World Impact

The main experiment contrasts the disclosure composite against six successive versions of external Net Impact estimates ($n = 3{,}210$ fund-period observations). Our claim is not that Upright estimates constitute perfect ground truth impact, but that massive disclosure-impact divergence remains highly informative even under noisy external estimation.

Disclosure and external impact percentiles correlate at just $\rho = 0.118$ ($p < 0.001$), indicating substantial disclosure-impact divergence. Funds differ by MAE $= 30.9$ percentile points on average, while the top-quintile hit rate is only marginally above chance (22.2% vs. 20%). Furthermore, 6.1% of all observations are extreme mismatches (top-quintile disclosure vs. bottom-quintile impact estimate), as visualized in Figure 3. We discuss the institutional drivers behind this asymmetry in Appendix D and provide examples for extreme cases in Table 12 in the Appendix.

Conditioning the residual δ on investment strategy reveals persistent asymmetries: Buyout funds systematically over-communicate by $+4.4$ percentile points ($n = 1{,}524$), whereas Venture funds under-communicate by -9.0 points ($n = 954$). The upper tail of δ remains stable across successive data releases, suggesting that disclosure-impact divergence is not purely transient noise (Appendix B.2).

4.2 Expert Validation of LLM Scoring

An expert scores the impact of 10 randomly selected funds across 2 quarters according to the scoring framework, resulting in 380 independently

scored metrics. To isolate the value of our rubric, we construct a baseline using our $\sim$200-term positive list, achieving $r = 0.42$. Second, on the same 380 cells, we compare scores of an expert rater, where the LLM achieves $r = 0.850$ ($p < 0.001$) and MAE $= 9.32/100$, confirming measurable reasoning lift over lexical methods. Rubric-tier agreement is substantial (quadratic-weighted $\kappa = 0.77$), with 91.6% of cells within one tier of the expert. The LLM is mildly more generous (mean bias $+6.03$), yet its self-reported confidence is well-calibrated ($\rho(\text{conf}, |\text{err}|) = -0.286$).

Cross-model rank-consistency. To test robustness across model families, we repeat the evaluation using Claude Opus 4.7 (Anthropic, 2024) on the same 380 expert-rated cells. Agreement with the expert remains strong ($\rho = 0.762$ for Claude vs. $\rho = 0.850$ for Gemini), while inter-model rank consistency reaches $\rho = 0.690$ ($p < 10^{-15}$). Since downstream analysis operates on percentile ranks within each cohort, stable relative ordering is more important than absolute score calibration. Differences in score magnitude across prompts and model families are well documented in recent LLM evaluation literature (Kathan et al., 2025; Salinas and Morstatter, 2024; Mizrahi et al., 2024).

4.3 Operational Stability Under Longitudinal Deployment

Because the pipeline is designed for continuous future-quarter monitoring, robustness to unseen firms is critical. We therefore evaluate the retrieval pipeline on a held-out test cohort of 134 firms independent from filter tuning. A qualitative audit on 20 firms (10 tune, 10 test; 1,455 kept URLs fully labeled, 1,281 dropped URLs sampled) yields an extrapolated recall of approximately 0.97. In the held-out evaluation, mean URL retention remains stable between the tuning and test cohorts (13.5% vs. 11.7%), and none of four distributional tests reject equality (all $p > 0.10$). Notably, stability holds despite the test firms containing more raw URLs (median 1,238 vs. 914). This generalization is the operational warranty for future quarterly updates of existing or new GPs.

5 Discussion

Our results demonstrate that longitudinal ESG disclosure monitoring is feasible despite the structural volatility of archival web data. By combining adaptive retrieval with temporal reconstruction, the pipeline transforms fragmented HTML and PDF snapshots into a continuous multi-quarter corpus suitable for scalable NLP analysis. Furthermore, the strong alignment between our schema-constrained LLM and domain experts ($r = 0.850$) confirms that modern models can reliably operationalize unstructured sustainability narratives, moving beyond traditional reliance on standardized reporting formats. The core value of this framework lies in it going beyond point-in-time anomaly detection, rather acting as a tool to analyze private markets as they face a phased influx of complex regulatory regimes (e.g., SFDR, CSRD).

Crucially, grounding these text-derived scores against external impact estimates exposes a substantial decoupling between corporate communication and portfolio substance ($\rho = 0.118$). This divergence is starkest at the extreme tail, where polished sustainability communication frequently masks weak real-world impact metrics (e.g., 98.8th percentile disclosure versus a 2.9th percentile actual impact estimate, as seen in Table 12 of the Appendix). Note that these large residuals (δ) should not be interpreted as definitive proof of deceptive intent, rather providing a reproducible, data-driven signal for identifying persistent communication asymmetries that warrant qualitative follow-up.

Operationalization and Release. For future research, we release a continuously updated companion platform[2] exposing model rationales and external impact percentiles for human auditing.

6 Conclusion

We presented a longitudinal NLP pipeline for evaluating corporate impact disclosures in the private markets. Using archival web reconstruction, we constructed a continuous 109,759-file corpus spanning 436 firms across 9 quarters. Rather than producing a static artifact, our open-source framework serves as a dynamic monitoring instrument, designed to scale both cross-sectionally to new firms and longitudinally to future reporting cycles. We show that an LLM can reliably operationalize heterogeneous disclosures at scale. By grounding these scores against external impact estimates, the framework moves beyond isolated text analysis toward continuous, scalable monitoring of disclosure-impact divergence over time.

[2]Web page remains hidden to preserve anonymity

Limitations

Single-rater expert validation. The $n = 380$ expert-alignment evaluation is conducted against a single domain expert rather than multiple raters with inter-annotator agreement statistics. The reported $r = 0.850$ should be interpreted as an upper bound on agreement that an independent rater might achieve; a multi-rater Cohen's κ or Krippendorff's α would more strictly bound the rubric's reliability and is the natural next methodological step.

Upright Net Impact Ratio as imperfect ground truth. The headline finding contrasts our disclosure scores against the Upright Project's Net Impact Ratio, but the latter is itself a modelled estimate over a proprietary taxonomy and is subject to its own measurement error. The $\rho = 0.118$ should be read as agreement between two noisy signals, not against a flawless ground truth. The monotonic rise of ρ across six successive Upright methodology releases ($0.07 \rightarrow 0.15$) is consistent with both signals becoming more reliable over time but cannot disentangle the two sources of improvement.

Wayback Machine coverage and recency bias. The Internet Archive's CDX index has uneven coverage across firms and time. Smaller and recently-launched GPs are systematically under-represented, and the mid-quarter snapshot rule cannot recover information that was never archived. The carry-forward imputation reduces but does not eliminate this bias; firms whose websites are rarely crawled may appear more static than they are in reality.

English-only LLM scoring. Although the URL filter recognizes multilingual ESG vocabulary (German *Nachhaltigkeit*, French *responsabilité*, etc.), the LLM scoring rubric and prompt are English-only. Non-English disclosure content is scored under the same anchored scale, which may systematically under-weight substantive disclosure that is neither translated nor paraphrased in English on the firm's site.

LLM scorer non-determinism and drift. Although we set `temperature` $= 0$ and enforce a JSON schema, modern LLM APIs are not strictly deterministic across inference calls, and the Gemini 3 Flash snapshot was updated during our study window. A small fraction of cells may have been scored under slightly different model versions. The released per-cell justifications allow any discrepancy to be detected post-hoc.

Heuristic URL filter and known failure modes. The filter is hand-tuned and exhibits documented failure modes (Appendix B.1), most notably a high-severity case where the strong-ESG override does not protect against the structural `NOISE_PATH_PATTERNS` loop in `should_keep`, costing $\sim$30 substantive false negatives across the audited sample. A one-line fix is staged for the v1 dataset release; numbers in this paper use the prefix filter to keep all reported metrics reproducible from the released code as of submission.

Per-category statistical power. The per-category Spearman correlations underlying the operationalizability gradient (Appendix B.2) are each estimated on $n = 20$ paired cells, giving a critical $\rho = 0.444$ for $p < 0.05$. The ranking at the extremes (*GHG*, *taxes* at the strong end; *meaning & joy* at the weak end) is therefore robust; the ordering within the middle of the gradient is noisier and should not be over-interpreted.

Causal interpretation of structured residuals. The Buyout/Venture residual asymmetry and the temporal rise of ρ are reported as structural patterns, not causal claims. The analyses in §4 and Appendix B.3 do not control for confounders including fund size, regulatory regime, year of website redesign, or LP composition. A formal regression treatment of these confounders is left to follow-up work.

Impact washing as inferred rather than observed. Throughout the paper, large positive residuals $\delta = \text{disclosure}_{\text{pct}} - \text{Upright}_{\text{pct}}$ are described as signals of potential impact washing. This framing is contingent on the assumption that truthful disclosure correlates strongly with measured impact, an assumption neither the Upright Net Impact Ratio nor any other public dataset can directly validate. We therefore recommend reading δ as a *misalignment* signal suitable for triage and qualitative follow-up, not as a binary impact washing classifier.

Ethical Considerations

Potential for Reputational Harm and Misuse. We present a computational pipeline capable of highlighting discrepancies between corporate sustainability claims and external impact estimates.

While this framework serves as a valuable triage tool for researchers, regulators, and investors, it carries a risk of dual-use if its outputs are weaponized or interpreted as definitive proof of deceptive intent (i.e., "greenwashing"). To mitigate unwarranted reputational harm to real-world financial entities, we explicitly frame the disclosure-impact divergence (δ) as a statistical residual that signals a need for human-in-the-loop qualitative follow-up, rather than a binary classification of corporate fraud. To further protect against direct reputational damage, we have anonymized the specific firm names in the extreme-divergence tails within this manuscript (e.g., Table 12), though the underlying code and methodology remain open-source to preserve scientific reproducibility.

Reliance on Modeled Ground Truth. Our framework relies on the Upright Project's Net Impact Ratio as an external baseline for real-world impact. It is ethically necessary to acknowledge that this external baseline is itself a modeled estimate subject to its own taxonomy, measurement errors, and biases. A high divergence score (δ) for a specific firm may not strictly indicate overcommunication; it could instead reflect a blind spot in the external impact model's coverage of that firm's specific niche or investment strategy. Users of this pipeline must contextualize the NLP-derived disclosure scores against the inherent limitations of the chosen comparative baseline.

Linguistic and Geographic Fairness. As noted in our limitations, the LLM scoring rubric and prompts used in our pipeline are exclusively in English. Although our URL filtering heuristics capture some multilingual web structures, substantive disclosures written in other languages that are neither translated nor paraphrased on a firm's English-language site may be systematically under-scored. Ethically, this introduces a geographic fairness issue, as the pipeline may inadvertently penalize non-English-speaking GPs and artificially inflate their divergence metrics. Operationalizing this tool in global regulatory contexts requires adapting the LLM evaluation to the native language of the reporting entities to ensure equitable assessment.

Data Provenance and Copyright. The longitudinal corpus introduced in this work relies on public web data reconstructed via the Internet Archive's Wayback Machine. We restrict our data collection strictly to publicly available corporate disclosures and reports, and our retrieval filters actively exclude non-relevant personal data (e.g., individual biographies are filtered out as structural noise, per Appendix A.1). In releasing our dataset and code, we adhere to fair use principles for academic text-mining, redistributing only cleaned text payloads and derived scores while respecting the source domains. Finally, to respect the intellectual property of our baseline provider, we do not redistribute the proprietary Upright Net Impact Ratio, publishing only the comparative statistics and residuals necessary to validate our NLP pipeline.

Use of LLMs. We disclose the use of LLMs for writing individual, standard pieces of code, as well as drafting and reviewing individual passages of this manuscript. Any LLM-generated output has been thoroughly checked by domain experts and we take full responsibility for all content presented herein.

References

Jefferson Abraham, Marcel Olbert, and Florin Vasvari. 2024. Esg disclosures in the private equity industry. *Journal of Accounting Research*, 62(5):1611–1660.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com/news/claude-3-family. Accessed: 2026-05-24.

Laura Braun and Christian Oswald. 2025. Travelwarncrawler: Constructing longitudinal datasets of government-issued travel warnings for political and social science research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 29–35.

Arion Cheong, Huijue Kelly Duan, Qing Huang, Miklos A Vasarhelyi, and Chanyuan Abigail Zhang. 2022. The rise of accounting: Making accounting information relevant again with exogenous data.

Marianne Chuang, Gabriel Chuang, Cheryl Chuang, and John Chuang. 2025. Judging it, washing it: Scoring and greenwashing corporate climate disclosures using large language models. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 17–31.

European Union. 2019. Regulation (EU) 2019/2088 of the european parliament and of the council of 27 november 2019 on sustainability-related disclosures in the financial services sector. Official Journal of the European Union, L 317/1.

European Union. 2022. Directive (EU) 2022/2464 of the european parliament and of the council of 14 december 2022 as regards corporate sustainability reporting. Official Journal of the European Union, L 322/15.

Rajna Gibson Brandon, Simon Glossner, Philipp Krueger, Pedro Matos, and Tom Steffen. 2022. Do responsible investors invest responsibly? *Review of Finance*, 26(6):1389–1432.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Manuel C Kathan, Sebastian Utz, Gregor Dorfleitner, Jens Eckberg, and Lea Chmel. 2025. What you see is not what you get: Esg scores and greenwashing risk. *Finance Research Letters*, 74:106710.

Christopher Marquis, Michael W Toffel, and Yanhua Zhou. 2016. Scrutiny, norms, and selective disclosure: A global study of greenwashing. *Organization science*, 27(2):483–504.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

The Upright Project. 2020. Quantifying the net impact of companies. Whitepaper, The Upright Project.

The Upright Project. 2026. The Upright Platform: Net Impact data. Data platform. Accessed: 2026-05-23.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

A Pipeline Implementation Details

A.1 Candidate URL filtering

We apply a deterministic two-sided path filter to every CDX URL. A positive list of $\sim$200 impact-relevant keywords is matched against the URL path. A blocklist of structural noise patterns (individual bios, press tombstones, locale duplicates, CMS attachment slugs) is then applied, with conflicts resolved by **positional precedence**: the first matching segment wins. A *strong-ESG override* rescues URLs whose path contains an unambiguous regulatory signal (e.g. `sfdr`, `article-9`, `tcfd`, `net-zero`) even if a blocklist segment appears earlier. Homepages and top-level portfolio overview pages are unconditionally retained for strategic context. Keyword and blocklist inventories are released verbatim in `shared.py`; the override regex is reproduced in Table 1.

Group	Representative tokens
Regulatory	`sfdr`, `csrd`, `tcfd`, `tnfd`, `eu-taxonomy`, `pai-statement`, `article-(8|9|10)`, `art-29`
Climate	`net-zero`, `decarbonisation`, `carbon-neutral`, `clean-energy`, `energy-transition`, `sbti`, `climate-action`
Reports	`sustainability-report`, `esg-report`, `integrated-report`, `annual-and-sustainability`, `impact-report`
Social	`modern-slavery`, `human-rights`, `gender-equality`, `inclusion-diversity`, `health-and-safety`, `occupational-health`
Nature	`biodiversity`, `circular-economy`, `nature-positive`, `deforestation`
Multilingual	`nachhaltigkeit`, `nachhaltig` (DE); responsibility (EN); responsabilité (FR)

Table 1: Token groups inside the *strong-ESG override* regex (`shared.STRONG_ESG_OVERRIDE`). The full keyword list comprises $\sim$200 entries grouped under positive (impact-relevant) and negative (structural noise) classes; both are published verbatim in the released code.

A.2 Mid-quarter snapshot selection

For each (canonical URL, quarter) pair we select the snapshot whose capture date d minimises $|d - m_q|$, where m_q is the mid-month of quarter q (Feb-15, May-15, Aug-15, Nov-15). An adaptive per-URL snapshot cap prevents heavily re-archived pages from dominating the corpus:

- ≤ 50 unique URLs in firm: no cap;
- 50-150: max 4 snapshots per URL;
- 150-300: max 3;
- >300: max 2.

A.3 Carry-forward state reconstruction

For each (firm, URL) we observe a set of capture quarters $\mathcal{Q}_u = \{q_1, \ldots, q_k\}$. We define the URL's *active interval* as $[q_{\min}, q_{\text{end}})$ where q_{end} is the earliest quarter satisfying one of:

- a 404 or 410 response is observed for that URL in the same CDX query window;
- no capture occurs for ≥ 3 consecutive quarters after the most recent capture.

For each quarter q in the active interval where no snapshot exists, we carry forward the cleaned text of the most recent prior snapshot (LOCF). A parallel track applies the same rule to discovered PDFs; PDFs without machine-readable text are OCR'd via Tesseract before inclusion.

Reproducibility. `reconstruct_quarterly.reconstruct_timeline`.

A.4 LLM scoring

Each reconstructed firm-quarter corpus is sent to `Gemini 3 Flash` (`temperature` $= 0$) under a single frozen system prompt that is released verbatim in the code. The prompt encodes:

1. a senior-analyst persona grounded in IMP, GIIN IRIS$^+$, UN SDG, and SFDR vocabulary;
2. the 19-category rubric across four pillars;
3. six anchored tier definitions (0-10, 11-25, 26-45, 46-65, 66-85, 86-100) with examples;
4. nine analysis guidelines (substance over keywords, specificity premium, framework alignment, etc.);
5. three mandatory reason codes for 0-scores (*not material*, *material but undisclosed*, *insufficient disclosure*).

A.5 Composite-score aggregation

Per-firm-quarter category outputs $\{s_c\}_{c=1}^{19}$ are aggregated post hoc with no further LLM calls. Let $\mathcal{A} = \{c : s_c > 0\}$ be the number of *addressed* categories, with $n_a = |\mathcal{A}|$. Let $n_r = 19 - |\{c : s_c =$

Field	Type
`fund_manager`	string
`analysis_date`	string
`society`	object of 5 CategoryScore
`knowledge`	object of 4 CategoryScore
`health`	object of 5 CategoryScore
`environment`	object of 5 CategoryScore
CategoryScore	
`score`	int $[0, 100]$
`confidence`	int $[0, 100]$
`explanation`	string (1-2 sentences)

Table 2: JSON schema enforced on every LLM response. Validation is performed via Pydantic before the result is persisted.

$0 \wedge \text{reason}_c = \text{not material}\}|$ be the set of *relevant* categories. We define

$$\text{depth} = \frac{1}{n_a} \sum_{c \in \mathcal{A}} s_c \qquad \text{breadth} = \frac{n_a}{n_r}$$

$$\text{composite} = \text{depth} \times \text{breadth}$$

Depth alone rewards specialists; breadth alone rewards generalists. Multiplying them penalizes both shallow-broad ("ESG bingo") and narrow-deep ("single-issue boutique") profiles.

B Extended Experimental Results

B.1 Filter stability

Table 3 reports the per-metric breakdown of URL retention across web pages, PDF documents, and the combined retention rate. No test rejects equality at any conventional threshold.

	Web	PDF	Total
n_{tune}	302	214	302
n_{test}	133	88	133
mean tune retention	13.2%	28.9%	13.5%
mean test retention	11.6%	32.6%	11.7%
median tune	9.7%	20.1%	10.3%
median test	8.6%	26.4%	9.3%
Δ mean (pp)	-1.6	$+3.7$	-1.8
95% bootstrap CI on Δ	$[-3.7, +0.7]$	$[-3.6, +11.3]$	$[-3.9, +0.4]$
Mann-Whitney p	0.286	0.670	0.274
Welch's t p	0.174	0.349	0.104
KS D	0.076	0.135	0.087
KS p	0.632	0.186	0.454
Permutation p (10k)	0.196	0.325	0.130
Cohen's d	-0.14	$+0.12$	-0.16
Rank-biserial r	$+0.06$	-0.03	$+0.07$
Test in tune IQR	50.4%	30.7%	51.9%

Table 3: Stability test across the three retention metrics. n_{tune} counts firms with at least one raw URL in that metric; tuning firms with zero PDFs are excluded from the PDF column, which explains its smaller n.

Site-size invariance. The held-out firms have larger archived websites than the tuning firms (median raw-CDX URLs 1,238 vs. 914, Mann-Whitney $p = 0.030$). Despite this, mean retention remains statistically indistinguishable, demonstrating that the filter is invariant not only to firm identity but also to site scale.

B.2 Expert validation

Full per-category breakdown. Table 4 reports the per-category agreement for all 19 rubric categories, sorted by Spearman ρ (descending). 18 of 19 are significant at $p < 0.05$; 17 at $p < 0.01$; 14 at $p < 0.001$.

Category	Pillar	n	ρ	r	MAE	Bias
GHG emissions	env.	20	0.920^{***}	0.917^{***}	7.0	$+5.5$
Taxes	soc.	20	0.916^{***}	0.888^{***}	4.0	-0.7
Biodiversity	env.	20	0.916^{***}	0.919^{***}	6.2	$+3.4$
Waste	env.	20	0.908^{***}	-	6.6	$+0.7$
Mental diseases	hth.	20	0.874^{***}	-	7.9	$+7.9$
Equality & human rights	soc.	20	0.861^{***}	0.839^{***}	8.9	$+2.7$
Societal stability	soc.	20	0.858^{***}	-	8.3	$+0.5$
Nutrition	hth.	20	0.856^{***}	-	10.4	$+10.4$
Physical diseases	hth.	20	0.837^{***}	-	10.4	$+6.8$
Jobs	soc.	20	0.819^{***}	-	9.3	$+7.0$
Scarce natural resources	env.	20	0.778^{***}	-	9.0	$+3.9$
Scarce human capital	knw.	20	0.774^{***}	-	11.2	$+6.1$
Non-GHG emissions	env.	20	0.761^{***}	-	10.0	$+8.8$
Knowledge infrastructure	knw.	20	0.712^{***}	0.631^{**}	13.4	$+10.3$
Relationships	hth.	20	0.656^{**}	-	7.1	$+3.1$
Creating knowledge	knw.	20	0.654^{**}	-	14.5	$+14.5$
Distributing knowledge	knw.	20	0.586^{**}	-	10.0	$+7.4$
Societal infrastructure	soc.	20	0.517^{*}	-	12.4	$+5.8$
Meaning & joy	hth.	20	0.375	0.353	11.0	$+10.9$

Table 4: All 19 categories.

Per-pillar agreement. Results are shown in Table 5

Pillar	n	ρ	MAE	Bias	κ_q
Society	100	0.865	8.6	$+3.1$	0.80
Knowledge	80	0.718	12.3	$+9.6$	0.53
Health	100	0.779	9.3	$+7.8$	0.67
Environment	100	0.897	7.7	$+4.4$	0.86

Table 5: Per-pillar agreement. Environment is strongest, Knowledge weakest.

Six-tier confusion matrix. Results are shown in Table 6

Materiality-flag consistency. For every 0-scored cell the LLM emits a reason code; the expert independently flagged cells as *not relevant*. Agreement on the binary materiality decision: TP $= 35$, FP $= 7$, FN $= 21$, TN $= 317$; precision 0.833, recall 0.625, $F_1 = 0.71$, accuracy 92.6%, Cohen's $\kappa = 0.67$.

expert\LLM	none	min.	mod.	sub.	comp.	best
none	57	41	18	0	0	0
minimal	2	56	30	6	1	0
moderate	2	11	46	25	4	0
substantial	0	1	10	45	9	0
comprehensive	0	0	0	2	11	2
best-in-class	0	0	0	0	1	0

Table 6: Confusion across the six anchored rubric tiers. 56.6% exact, 91.6% within-one-tier agreement. Quadratic-weighted $\kappa = 0.77$.

Binary disclosure detection at three thresholds. Results are shown in Table 7.

Threshold	Prec.	Recall	F_1	Acc.
any disclosure (> 0)	0.891	0.960	0.924	0.876
moderate or above (≥ 26)	0.738	0.917	0.818	0.818
substantial+ (≥ 46)	0.660	0.864	0.749	0.876

Table 7: LLM's per-cell decision treated as a binary classifier. Recall dominates precision at every threshold, consistent with the LLM's mild upward bias (+6.03 on the 0-100 scale).

Confidence calibration. The LLM's self-reported confidence is meaningfully predictive of its own error. Spearman $\rho(\text{confidence}, |\text{error}|) = -0.286$, and MAE drops monotonically across confidence tertiles are reported in Table 8

Conf. tertile	n	mean conf.	MAE
low	162	81.9	11.09
mid	108	90.0	10.02
high	110	95.8	6.04

Table 8: Confidence-tertile calibration.

Cross-quarter stability. Agreement is stable across the two evaluated quarters: 2024 Q3: $n = 190$, $\rho = 0.826$, MAE $= 10.07$, bias $+8.01$, $\kappa_q = 0.72$; 2025 Q4: $n = 190$, $\rho = 0.838$, MAE $= 8.57$, bias $+4.05$, $\kappa_q = 0.81$.

B.3 Disclosure vs. Upright

Per-period breakdown. Results are shown in Table 9

Stratified comparisons (pooled across periods). Results are shown in Table 10.

Quintile transition matrix. Results are shown in Table 11.

Upright vers.	n	ρ	p	MAE	Top-Q	Bot-Q
1.6.0 (09/24)	535	0.071	0.100	32.2	0.260	0.220
1.7.0 (11/24)	535	0.090	0.037^{*}	31.7	0.223	0.275
1.8.0 (04/25)	535	0.121	0.005^{**}	30.7	0.194	0.296
1.9.0 (06/25)	535	0.130	0.003^{**}	30.4	0.178	0.287
1.10.0 (10/25)	535	0.145	$<0.001^{***}$	30.1	0.240	0.296
1.11.0 (12/25)	535	0.149	$<0.001^{***}$	30.0	0.238	0.301

Table 9: Per-release agreement. Top-Q / Bot-Q are top-quintile and bottom-quintile hit rates (chance baseline 0.20).

Stratum	n	ρ	MAE	Bias
strategy				
Buyout	1,524	0.090^{***}	32.6	+4.4
Growth	576	0.161^{***}	30.9	+6.4
Venture	954	0.143^{***}	28.7	−9.0
Other	156	0.213^{**}	26.3	−11.0
region				
EU/EEA	2,208	0.113^{***}	31.8	+2.1
non-EU	1,002	0.131^{***}	28.8	−4.7
vintage				
2015-2019	1,314	0.145^{***}	30.1	+1.0
2020+	1,896	0.097^{***}	31.4	−0.7

Table 10: Strategy is the strongest source of bias asymmetry; region and vintage are second-order.

GP-level robustness. Aggregating fund percentiles to the GP level (mean of fund percentiles within each GP) yields $n = 2{,}442$ (GP $\times$ period) observations, Spearman $\rho = 0.141$ ($p < 0.001$), MAE 30.10, top-quintile hit rate 23.4%, bottom-quintile 32.2%. The result is essentially unchanged from the fund-level analysis, ruling out a multi-fund-per-GP artifact.

Most extreme over- and under-communicators. Results are shown in Table 12.

C Released Dataset

C.1 Corpus statistics

Results are shown in Table 13.

C.2 Panel composition

The 535 funds backing the 407 unique GPs are distributed as follows: Buyout 254, Venture 159, Growth 96, Other 26. 2,208 of 3,210 fund-period observations (68.8%) are from EU/EEA-domiciled funds.

C.3 Provenance and release format

For every quarter-state file the release ships the original Wayback timestamp, the canonical URL, the carry-forward source quarter, and the cleaned-text length. PDF text additionally carries an `ocr_used`

Upr.\Disc.	Q1	Q2	Q3	Q4	Q5
Q1 (bot.)	185	126	128	109	**114**
Q2	154	152	109	132	129
Q3	99	142	133	118	142
Q4	121	120	130	137	108
Q5 (top)	**83**	105	142	154	138

Table 11: Quintile transition (pooled, rows = Upright, cols = disclosure). Bold cells = extreme misclassifications (top×bottom). Exact-quintile agreement 23.2%; within-one 55.6%.

GP	Strategy	Disc. pct	Upr. pct	δ
Over-communicators (disc. ≫ Upright)				
Fund A	Growth	98.8	2.9	+95.9
Fund B	Venture	99.3	4.5	+94.9
Fund C	Buyout	99.9	5.0	+94.9
Fund D	Growth	96.6	3.4	+93.3
Under-communicators (Upright ≫ disc.)				
Fund E	Venture	1.6	97.8	−96.2
Fund F	Venture	4.5	99.6	−95.1
Fund G	Venture	8.2	99.9	−91.7

Table 12: Highest-magnitude residuals across the six Upright releases. Each named GP appears in this tail across *multiple* releases, indicating a stable rather than transient mismatch.

flag. The companion site (anonymised) exposes the per-cell LLM justification alongside the Upright percentile and the signed residual δ, so that any third party can re-score under an alternative rubric, LLM, or scoring methodology. **Note:** we release *our* scoring outputs, but the Upright Net Impact Ratio remains the proprietary data of The Upright Project and is not redistributed; we publish only our comparison statistics against it.

D Extended Financial Discussion

D.1 Interpreting the Buyout/Venture asymmetry

The +4.4 pp Buyout / −9.0 pp Venture residual asymmetry (§4.1) is consistent with three institutional drivers visible in the underlying funds. First, Buyout funds raise from large institutional LPs (pension funds, sovereign wealth) whose investment committees increasingly mandate explicit ESG disclosure in the side-letter; this creates direct economic pressure to communicate, independent of the substantive impact of the underlying portfolio. Second, Buyout GPs at scale maintain dedicated communications and IR functions, providing the staff bandwidth to produce framework-conformant disclosures (TCFD, SBTi, CSRD-style reports). Third, the impact of Venture funds is typically embedded in the product roadmap of their portfolio companies (e.g. a climate-tech start-up displacing fossil-fuel infrastructure) rather than in ongoing portfolio-management practices the GP would naturally narrate. The Venture GP captures the upside in the cap table; the impact lives in the start-up's customer base.

Stage	Count
GPs in input cohort	440
GPs with reconstructable presence	436
Calendar quarters covered	9 (2024 Q1-2026 Q1)
Raw CDX URLs queried	1,400,413
HTML	1,349,374
PDF	51,039
Candidate URLs (post-filter)	71,672
HTML	64,929
PDF	6,743
Unique HTML pages fetched	∼22,600
Unique PDF reports fetched	∼2,300
Unique source documents	**∼ 25,000**
Quarter-state files (carry-forward)	109,759
(GP × quarter) panel cells	3,718
Cleaned text payload	≈ 0.95 GB
Raw HTML + PDF (where licensed)	≈ 17 GB

Table 13: Corpus volumes through the pipeline.

D.2 Regulatory context

The European PE market faces three converging disclosure regimes during our 2024-2026 study window: **SFDR** (Sustainable Finance Disclosure Regulation), with its Article 8 / 9 classification and Principal Adverse Impact (PAI) statements; **CSRD** (Corporate Sustainability Reporting Directive), phasing in from FY2024 reporting; and the **EU Taxonomy** for substantial-contribution and do-no-significant-harm assessments. The legal sufficiency of GP disclosure is therefore in active flux: a fund with a thin website today may be obliged to publish substantially more by the next reporting cycle. Our quarterly panel is the temporal granularity needed to detect the resulting transition dynamics, which annual-only analyses (e.g. 10-K-style or yearly CSR-report studies) miss by construction.

D.3 Use cases

Three concrete audiences emerge from our results. **Limited Partners** performing pre-commitment due diligence can use the per-firm δ as a triangulation against the GP's self-reported impact narrative, reducing reliance on self-administered DDQ responses. **Regulators** can use the panel to monitor cohort-level changes in disclosure depth and

breadth in response to SFDR/CSRD implementation milestones, and to identify firms whose disclosure diverges sharply from independent impact estimates. **Investigative journalists and academics** can use the named-GP outlier tail (App. B.3) as a sampling frame for qualitative greenwashing case studies, with full source-text provenance attached to every claim.

D.4 Comparison to commercial ESG ratings

Commercial ESG rating providers (MSCI ESG, Sustainalytics, S&P Global, Refinitiv) operate predominantly on public-equity issuers, treat PE GPs only obliquely, and publish neither their input documents nor their per-criterion scores. The closest public-data analogue is the Upright Project's Net Impact Ratio, which we use as our external benchmark. Our pipeline is complementary rather than competitive with these providers: we do not estimate *impact*; we measure the *depth and breadth of disclosure* and surface its mismatch with a third-party impact estimate. The correlation $\rho = 0.118$ should be read not as a weakness of either signal but as evidence that what GPs *say* is, at best, a noisy proxy for what they *do*; the empirical claim that motivates the paper.