GRAB: A Risk Taxonomy–Grounded Benchmark for Unsupervised Topic Discovery in Financial Disclosures

Ying Li
The University of Edinburgh, UK
sunnie.y.li@ed.ac.uk

Tiejun Ma
The University of Edinburgh, UK
tiejun.ma@ed.ac.uk

Abstract

Risk categorization in 10-K risk disclosures matters for oversight and investment, yet no public benchmark evaluates *unsupervised* topic models for this task. We present **GRAB**, a finance-specific benchmark with **1.61M** sentences from **8,247** filings and *span-grounded* sentence labels produced without manual annotation by combining FinBERT token attention, YAKE keyphrase signals, and taxonomy-aware collocation matching. Labels are anchored in a risk taxonomy mapping **193** terms to 21 fine-grained types nested under five macro classes; the 21 types guide weak supervision, while evaluation is reported at the macro level. GRAB unifies evaluation with fixed dataset splits and robust metrics—*Accuracy*, *Macro-F1*, *Topic BERTScore*, and the entropy-based *Effective Number of Topics*. The dataset, labels, and code enable reproducible, standardized comparison across classical, embedding-based, neural, and hybrid topic models on financial disclosures.

1 Introduction

The SEC requires public companies to disclose material risk factors in a dedicated section of their annual 10-K filings. These *Item 1A: Risk Factors* sections contain information that materially affects investment decisions—changes in their content are associated with future stock return volatility, and investors react to their filing Campbell et al. [2014]. Yet extracting structured insight from these lengthy, legalistic texts remains difficult, in part because there is no public evaluation framework for how well *unsupervised* models recover financially meaningful *risk categories* from raw disclosures.

Prior work underscores the need for a risk-aware benchmark on regulatory text. Beyond surface variables such as disclosure volume, tone, and readability, the *risk category* (e.g., litigation, natural disasters) is central for analysis and decision-making, but it has received comparatively less attention in scalable evaluations Bao and Datta [2014], Huang and Li [2011], Campbell et al. [2014]. Manual coding does not scale: Mirakur reports hand-labeling 29 risk categories for 122 firms—far below 1% of annual 10-K filings—illustrating the impracticality of exhaustive human annotation Mirakur [2011]. Methodologically, the landscape spans classical probabilistic models such as LDA Blei et al. [2003], embedding-space extensions such as Gaussian LDA (GLDA) Das et al. [2015], neural contextualized approaches including CTM Bianchi et al. [2021], sentence-aware methods such as SentenceLDA Cha and Lee [2024] and SenClu Schneider [2024], and embedding-based clustering pipelines such as BERTopic Grootendorst [2022]. However, these advances have largely been assessed on general-domain corpora rather than against finance-specific risk taxonomies or the peculiarities of 10-K prose (boilerplate, context shifts, multiword financial phrases), and access to labeled financial text remains limited Jørgensen et al. [2023]. Notably, Sent-LDA Bao and Datta [2014] relied on a small, non-public labeled set, highlighting the need for an open benchmark.

We introduce **GRAB** (Grounded Risk-Aware Benchmark), a public benchmark for evaluating topic models on financial risk categorization without manual annotation. GRAB anchors labels in the

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Generative AI in Finance.

Hofeditz taxonomy Bender et al. [2016]: we map **193** curated risk terms into **21** fine-grained subcategories nested under the taxonomy's **five** macro risk classes. We use this structure for weak supervision, yielding span-grounded labels at scale via a blend of finance-aware attention, document-local keyphrase cues, and a curated lexicon. For evaluation, we operate at the level of the five macro risk classes, scoring how well models recover these categories.

GRAB also provides an evaluation protocol centered on risk categories. We report *Accuracy* and *Macro-F1* using a simple logistic classifier on sentence–topic mixtures, and *Topic BERTScore* Zhang et al. [2020] to assess semantic tightness. We do not report perplexity because it is not comparable across our baselines (e.g., discrete LDA vs. continuous GLDA Das et al. [2015]). Instead, we report the *Effective Number of Topics*—the exponential of Shannon entropy Shannon [1948]—which summarizes how concentrated or diffuse a sentence's topic mixture is (values near 1 indicate decisive assignments; values near K indicate diffuse mixtures), following the standard Hill-numbers formulation Hill [1973]. For category-level reporting, we evaluate against the five macro risk classes in a multilabel setting. In sum, GRAB provides both a finance-specific corpus and a clear protocol to test whether models recover sentence-level *disclosure risks*, a need underscored by evidence that sentence-level risk categories matter for markets Bao and Datta [2014].[1]

2 Benchmark Design

2.1 Dataset Preprocessing

We extract *Item 1A: Risk Factors* from annual 10-K filings of S&P 500 firms, yielding a corpus of **1,613,837** sentences across **2001–2025**. We apply light normalization to handle legal formatting and sentence boundaries (Appendix B), then segment into sentences. The scale and heterogeneity of this corpus capture diverse, domain-specific risk expressions.

2.2 Word Importance Scoring

We estimate per-token importance by blending *finance-tuned attention* with a *YAKE*-based phrase signal. Risk categories follow the Hofeditz taxonomy Bender et al. [2016], flattened from 193 curated terms into 21 subcategories $\mathcal{Y}$ and later used for risk-aware enhancement and evaluation mapping.

(i) Finance-tuned attention. FinBERT Huang et al. [2023] processes each sentence s and yields attention maps. We aggregate across layers/heads and take the CLS $\rightarrow$ token row as an attention score $A_i \in [0, 1]$ for token w_i, merging wordpieces by a max-over-pieces rule.

(ii) YAKE-based contextual importance. We run YAKE Campos et al. [2020] per sentence to obtain keyphrases p_j with raw scores $r_j > 0$ (lower is better), then convert them to token-level importance via: (1) per-sentence min–max normalization and inversion; (2) span $\rightarrow$ token assignment using boundary-aware, space/hyphen–tolerant matching; and (3) blending with attention and per-sentence max-normalization (details in Appendix A, hyperparameters in Appendix D):

$$I_i \;=\; \lambda\, A_i \;+\; (1-\lambda)\, Y_i, \qquad \tilde{I}_i \;=\; \frac{I_i}{\max_k I_k} \in [0,1].$$

We use $\lambda = 0.8$ in the main comparisons.

2.3 Risk-Aware Score Enhancement

To sharpen signal without labels, we apply a lightweight, precision-first enhancer *within* each sentence: **(i) Unigram boost**—upweight exact (case-insensitive) matches from the 21-subcategory lexicon (the fine-grained level under the five macro classes) by a fixed multiplier, capped at 1.0; **(ii) Collocation boost**—detect multiword financial terms with a boundary-aware, space/hyphen–tolerant, parenthesis-aware matcher and scale only the tokens inside the matched span (no sentence-wide multipliers; no lemmatization/singularization). The phrase list is the union of taxonomy-derived expressions and a curated finance dictionary; optional fuzzy tolerance is supported but off by default (Appendix D).

[1] We release corpus-construction scripts, weak labels, taxonomy anchors, collocation lists, and evaluation code at `https://github.com/Sunnie-Li/GRAB-Benchmark`.

2.4 Weak Label Generation

From each sentence, we take the top-m tokens by enhanced importance $\tilde{I}_i$ and accumulate evidence at the *subcategory* level (21 types) via three sources: **(i) lexicon**—exact, case-insensitive matches to the curated subcategory terms; **(ii) collocation**—boundary-aware matches of multiword financial phrases, recorded token-locally over the matched span; **(iii) semantic backoff**—for remaining high-importance tokens without a direct match, an optional lightweight kNN over subcategory terms in embedding space (disabled in the main configuration; Appendix D). Aggregating these signals yields a 21-dimensional per-sentence vector used only for evaluation. For reporting at the practitioner level, we deterministically roll up the 21 subcategories to their five parent macro classes.

2.5 Evaluation Protocol

We assess models along three aspects aligned with downstream use and topic quality. Let K denote the number of learned topics (default $K{=}21$; details in Appendix D):

- **Predictive utility:** train a one-vs-rest logistic regression on each model's sentence–topic mixtures to predict the *five* macro risk classes (the 21 subcategories are used only to construct weak labels and for fine-grained analysis, not as prediction targets); report *Accuracy* and *Macro-F1*[2].
- **Topic quality:** compute *Topic BERTScore* Zhang et al. [2020] by comparing, for each topic, its representative sentences to its top member sentences (sampling details in Appendix F).
- **Assignment decisiveness (descriptive):** report the *Effective Number of Topics* per sentence,

$$H(\theta) = -\sum_{k=1}^{K} \theta_k \log \theta_k, \qquad N_{\text{eff}}(s) = \exp\big(H(\theta)\big) \in [1, K],$$

  and summarize the mean$\pm$std on the test split. Lower values indicate more decisive (peaked) mixtures; values near K indicate diffuse mixtures. We use natural logs and the convention $0 \log 0 := 0$. How θ is obtained per model is detailed in Appendix E.

3 Experiments

3.1 Dataset Statistics

GRAB comprises 1,613,837 sentences from *Item 1A: Risk Factors* in 8,247 annual 10-K disclosures spanning 2001–2025 After light normalization (Appendix B), we segment into sentences; the corpus spans long-form legal prose and finance-specific multiword expressions (e.g., *cash flow*). To avoid look-ahead, we adopt a fixed *chronological split*[3] in which each sentence inherits its document's SEC release (filing) year: **Train** = $\leq$2022, **Dev** = 2023, **Test** = 2024–2025. We release sentence identifiers and filing metadata for reproducibility and per-company analysis; Appendix G includes a per-category prevalence chart (Fig. 2) and the disclosure/sentence distributions by year (Figs. 3, 4).

3.2 Baselines

We benchmark representative families—classical probabilistic, embedding-space, neural, sentence-level, and embedding–clustering with sparse term weighting:

- **LDA** Blei et al. [2003]: classical word-level topic model trained on BoW/TF–IDF.
- **GaussianLDA (GLDA)** Das et al. [2015]: topics are Gaussians in a word-embedding space (300d Word2Vec, GoogleNews Mikolov et al. [2013]); sentence-level mixtures are formed from within-sentence word–topic counts.
- **CTM** Srivastava and Sutton [2017], Bianchi et al. [2021]: a ProdLDA-style neural topic model that conditions BoW on contextual embeddings (Sentence-BERT Reimers and Gurevych [2019]).

[2]We omit Micro-F1: it aggregates over instances and is dominated by frequent classes, so it closely tracks Accuracy. Macro-F1 averages F1 across classes, giving equal weight to minority risk types and better capturing long-tail performance.

[3]By sentence count in the current release: $\sim$79.07% Train, 7.09% Dev, 13.84% Test. See sentence-count and disclosure-count distributions and plots in Appx. G.

Table 1: Topic discovery on GRAB (Item 1A sentences). ↑ higher is better.

Method	Accuracy ↑	Macro-F1 ↑	Topic BERTScore ↑	Eff. # Topics[†]
LDA (BoW/TF–IDF)	0.37	0.36	**0.84**	3.52
GLDA (Word2Vec)	0.35	0.23	**0.84**	1.30
CTM (BoW + SBERT)	**0.38**	**0.41**	**0.84**	5.92
SenClu (SBERT)	0.24	0.18	0.82	1.00
SentenceLDA (SLDA)	0.36	0.32	**0.84**	1.00
BERTopic	0.35	0.22	0.82	1.00
BERTopic-Soft	0.35	0.25	0.82	3.09

[†] Effective Number of Topics is a descriptive quantity (not an objective with a "higher/lower is better" direction). It takes values in $[1, K]$, where K is the number of topics: N_{eff}=1 denotes maximally decisive, single-topic assignments, while values approaching K indicate diffuse, near-uniform mixtures (see Sec. 2.5).

- **SenClu** Schneider [2024]: sentence-level scoring via similarity between a sentence embedding and learned topic prototypes (Sentence-BERT).
- **BERTopic** Grootendorst [2022]: embedding-based clustering with topic word weights from class-based TF–IDF (SBERT embeddings → UMAP → HDBSCAN/k-means → c-TF–IDF). We evaluate two modes: (i) *sentence-level* BERTopic (splitting Item 1A into sentences) and (ii) *document-level* **BERTopic-Soft** using HDBSCAN soft assignments (`calculate_probabilities=True`) to obtain topic mixtures without splitting.

Unless noted, we fix the number of topics K across methods for comparability and align discovered topics to the *five* macro risk classes (Sec. 2). For all systems we derive soft $\theta(s)$ as described in Appx. E; full hyperparameters are in Appx. D.

3.3 Results

Table 1 summarizes performance on GRAB. *CTM* attains the best label-aware scores (Macro-F1 $= 0.41$, Accuracy $= 0.38$), with *LDA* close behind ($0.36/0.37$). Clustering- or sentence-centric methods are lower on these metrics (*SenClu*: $0.18/0.24$; *BERTopic*: $0.22/0.35$), while *SLDA* is competitive ($0.32/0.36$). *Topic BERTScore* is uniformly high and tightly grouped (0.82–0.84), with CTM, LDA/GLDA, and SLDA at the upper end (0.84). The *Effective Number of Topics* N_{eff} reflects assignment concentration: hard-assignment pipelines yield values near 1 (SenClu, BERTopic, SLDA), the soft BERTopic variant increases it (3.09), LDA/GLDA sit in between ($3.52/1.30$), and CTM produces the broadest mixtures (5.92). Qualitatively, we observe that (i) boilerplate language induces spurious topics that dilute category signals, (ii) polysemous tokens (e.g., *short*, *term*) shift meaning across market/credit/accounting contexts, and (iii) long-tail but material categories remain under-recovered by frequency-driven approaches. Per-category results and sensitivity to K and enhancer settings are in Appx. D; aggregate risk-count and prevalence plots appear in Appx. G.

4 Conclusion

We present GRAB, a public benchmark for *unsupervised* risk categorization in 10-K Item 1A risk-factor sections. GRAB replaces traditional manual labels with financial-term–grounded weak supervision to derive a taxonomy-driven prior, yielding sentence-level labels across S&P,500 disclosures. Labels are anchored in a fine-grained risk taxonomy: subtypes guide weak supervision, and evaluation is reported at the macro level. Using fixed chronological splits and a risk-aware protocol—*Accuracy*, *Macro-F1*, *Topic BERTScore*, and the *Effective Number of Topics*—the benchmark enables fair, reproducible, and robust comparison across classical, embedding-based, neural, and hybrid topic models on standardized financial risk disclosures. Empirically, GRAB shows that contextualized methods improve label-aware performance while traditional word-based models remain competitive, and that coherence-style signals alone do not capture risk-type recovery, underscoring the need for domain-aware evaluation. This work supports transparent, unified and robust benchmarking and follow-on research on risk extraction; our results point to future directions including better-calibrated sentence-level mixtures aligned with risk taxonomies, robustness to boilerplate and context shifts, and evaluation that bridges discovered risk categories to market conditions and compliance outcomes.

References

Yang Bao and Anindya Datta. Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, 60(6):1371–1391, 2014. URL http://www.jstor.org/stable/42919610.

Micha Bender, Sven Panz, and Dietmar Hofeditz. A general framework for the identification and categorization of risks: An application to the context of financial markets. *SSRN Electronic Journal*, 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3738273. SSRN: 3738273.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.96. URL https://aclanthology.org/2021.acl-short.96.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

John L Campbell, Hsinchun Chen, Dan S Dhaliwal, Hsin-min Lu, and Logan B Steele. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1):396–455, 2014.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

Taehun Cha and Donghun Lee. Sentencelda: Discriminative and robust document representation with sentence level topic model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 521–538, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.31. URL https://aclanthology.org/2024.eacl-long.31/.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian LDA for topic models with word embeddings. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1077. URL https://aclanthology.org/P15-1077/.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

Mark O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973. doi: 10.2307/1934352.

Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

Ke-Wei Huang and Zhuolun Li. A multilabel text classification algorithm for labeling risk factors in sec form 10-k. *ACM Transactions on Management Information Systems (TMIS)*, 2(3):1–19, 2011.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. MultiFin: A dataset for multilingual financial NLP. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.66. URL https://aclanthology.org/2023.findings-eacl.66/.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Yuka Mirakur. Risk disclosure in SEC corporate filings. Working paper, University of Pennsylvania, 2011. URL http://repository.upenn.edu/wharton_research_scholars/85. Accessed October 1, 2013.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Johannes Schneider. Topic modeling with fine-tuning llms and bag of sentences, 2024. URL https://arxiv.org/abs/2408.03099.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=H1TMjf9xx.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Looks solid! I made a few small fixes to line everything up with the five-macro-class evaluation and to clean minor issues (added the explicit "five macro classes" target in the frozen-knobs bullet, fixed a missing file extension, and ensured no leftover "Sharpness" references remain). Here's the updated appendix block you can drop in:

A YAKE-based Context Equations

For each sentence s, YAKE Campos et al. [2020] returns keyphrases p_j with raw scores $r_j > 0$ (lower is better). We convert these to token-level *importance* via:

$$\hat{y}_j = 1 \; - \; \frac{r_j - r_{\min}}{r_{\max} - r_{\min} + \varepsilon} \in [0,1], \qquad \text{(per-sentence min–max normalization \& inversion)} \tag{1}$$

$$Y_i = \max_{j:\, w_i \in p_j} \hat{y}_j \;\; \text{(or 0 if no phrase covers } w_i), \qquad \text{(span} \rightarrow \text{token assignment)} \tag{2}$$

$$I_i = \lambda\, A_i + (1-\lambda)\, Y_i, \qquad \text{(blend attention with YAKE)} \tag{3}$$

$$\tilde{I}_i = \frac{I_i}{\max_k I_k} \in [0,1]. \qquad \text{(per-sentence max-normalization)} \tag{4}$$

Here A_i is the FinBERT CLS$\rightarrow$token attention for w_i, $\lambda \in [0,1]$ is the blend weight, and $\varepsilon > 0$ is a small constant.

B Preprocessing Rules

- Remove control characters: `[\x00-\x1f\x7f-\x9f]`.
- Collapse repeated whitespace.
- Replace `\bNo\.` with `Number` (word-boundary anchored).
- Join consecutive lines if the trimmed last character $\notin \{.,?,!\}$.

C Matching Rules (Risk-Aware Enhancement)

Unigrams (single-word). Case-insensitive exact match against the taxonomy's single-word tokens. For any matched token w:

$$\tilde{I}_w \leftarrow \min\{1.0,\; \beta_{\text{uni}} \cdot \tilde{I}_w\}.$$

Collocations (multi-word). Boundary-aware, case-insensitive, hyphen/space–tolerant patterns built from (i) taxonomy phrases and (ii) a curated finance dictionary (union). Examples:

- `A-Shares` ↔ `A Shares`
- `Accredited Asset Management Specialist (AAMS)` ↔ `AAMS`

If a phrase matches over token span S, apply a *token-local* boost:

$$\forall\, w \in S: \quad \tilde{I}_w \leftarrow \min\{1.0,\ \beta_{\text{col}} \cdot \tilde{I}_w\}.$$

(Parenthetical abbreviations yield both long-form and acronym patterns. Fuzzy tolerance is *off* by default.)

D Default Settings

- Attention/context blend $\lambda = 0.8$ in (3).
- Tokens kept per sentence for labeling: $m = 10$.
- Boosts/cap: $\beta_{\text{uni}} = 1.5$, $\beta_{\text{col}} = 1.2$, cap $= 1.0$.
- YAKE per-sentence normalization uses (1) with $\varepsilon \approx 10^{-9}$.
- Collocation fuzzy tolerance: off by default.

E Effective Number of Topics: Definition and Implementation

Definition. Given a per-sentence topic mixture $\theta(s) \in \Delta^{K-1}$ over K topics, we quantify assignment concentration via the entropy-derived *Effective Number of Topics* (Hill number of order 1):

$$H(\theta) \;=\; -\sum_{k=1}^{K} \theta_k \log \theta_k, \qquad N_{\text{eff}}(s) \;=\; \exp\big(H(\theta)\big) \in [1, K].$$

Lower values indicate more decisive (peaked) mixtures; values near K indicate diffuse mixtures. We use natural logs and the convention $0 \log 0 := 0$.

Computation notes. We L1-normalize θ before computing $H(\theta)$. For numerical stability, add a small ε inside the log if needed. We summarize N_{eff} by mean±std on the test split. This statistic is descriptive and complements label-aware and semantic metrics.

Obtaining $\theta(s)$ per model. To ensure comparability across methods, we compute a *soft* topic mixture $\theta(s)$ for each sentence s as follows (with a small $\epsilon > 0$ for numerical stability and L1-normalization in all cases):

- **LDA** Blei et al. [2003]: use the inferred per-document topic proportions. If only word–topic counts $n_{s,k}$ are available, set $\theta_k(s) \propto n_{s,k} + \epsilon$.
- **GLDA** Das et al. [2015]: aggregate word–topic assignments within the sentence to counts $n_{s,k}$, then set $\theta_k(s) \propto n_{s,k} + \epsilon$.
- **CTM / ProdLDA family** Srivastava and Sutton [2017], Bianchi et al. [2021]: take the variational mean logits μ_s and set $\theta(s) = \mathrm{softmax}(\mu_s)$.
- **SenClu** Schneider [2024] **and BERTopic** Grootendorst [2022]: compute $\theta_k(s) \propto \exp\big(\cos(e(s), c_k)/\tau\big)$ from a sentence embedding $e(s)$ and topic centroid c_k (fixed temperature τ); HDBSCAN outliers in BERTopic are assigned to the nearest centroid before the softmax.
- **SentenceLDA (SLDA)** Cha and Lee [2024]: one topic per sentence; use a one-hot $\theta(s)$.

F Split Protocol and Reproducibility

Chronological split. We assign each document to a year by its SEC *release (filing) date*, and each sentence inherits its document's year. We use a fixed chronological split: **Train** = up to and including 2022, **Dev** = 2023, **Test** = 2024–2025. The corpus spans 2001–2025 (8,247 Item 1A disclosures) with a total of 1,613,837 sentences. Sentence-level split sizes are:

$$\text{Train } 1{,}276{,}105\ (79.07\%), \qquad \text{Dev } 114{,}422\ (7.09\%), \qquad \text{Test } 223{,}310\ (13.84\%).$$

Year-by-year disclosure and sentence distributions are shown in Appendix G.

Table 2: Chronological split by *release year* with sentence counts.

Split (Years)	# Sentences	Split Percentage
Train ($\leq$2022)	1,276,105	79.07%
Dev (2023)	114,422	7.09%
Test (2024–2025)	223,310	13.84%

Frozen evaluation knobs.

- Topics: default $K{=}21$; report mean$\pm$std over *S* seeds (placeholders).
- Predictive utility: logistic regression (one-vs-rest) with fixed seed and default regularization; identical features across methods; *targets are the five macro risk classes*; report **Accuracy** and **Macro-F1** on Dev/Test.
- Topic BERTScore: `roberta-large`, IDF on, baseline rescaling on; per-topic representative selection and top-member sampling as in the main text.
- Effective Number of Topics: computed per sentence from θ (Appendix E); hard-assign models treated as one-hot by default; report mean$\pm$std on Test.
- Weak labels: top-m tokens per sentence; identical across systems.

Leakage checks. (1) Each 10-K filing is assigned to exactly one split by SEC release year (no document appears in multiple splits); (2) after normalization, identical sentence strings are deduplicated so that no sentence occurs in more than one split. Detected violations are logged and removed prior to release.

G Data Visualization

To contextualize label balance and temporal coverage, we provide four summary figures: (i) a bar chart of *macro* category prevalence aggregated across issuers (Fig. 1); (ii) a bar chart of the 21 *subcategory* labels (Fig. 2); (iii) yearly counts of Item 1A *disclosures* by SEC release year (Fig. 3); and (iv) yearly counts of *sentences* by SEC release year (Fig. 4). The two timeline plots are shaded with the fixed split bands (Train $\leq$2022, Dev $=$ 2023, Test $=$ 2024–2025).

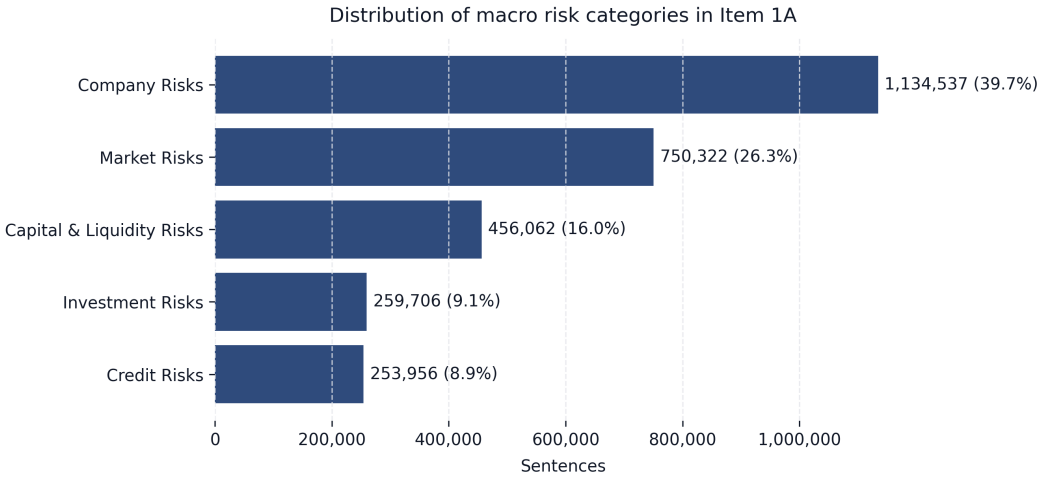

Figure 1: Distribution of weakly-labeled risk sentences by *macro* category (S&P 500, 2001–2025). Bars show counts; labels indicate count and share of all labels.

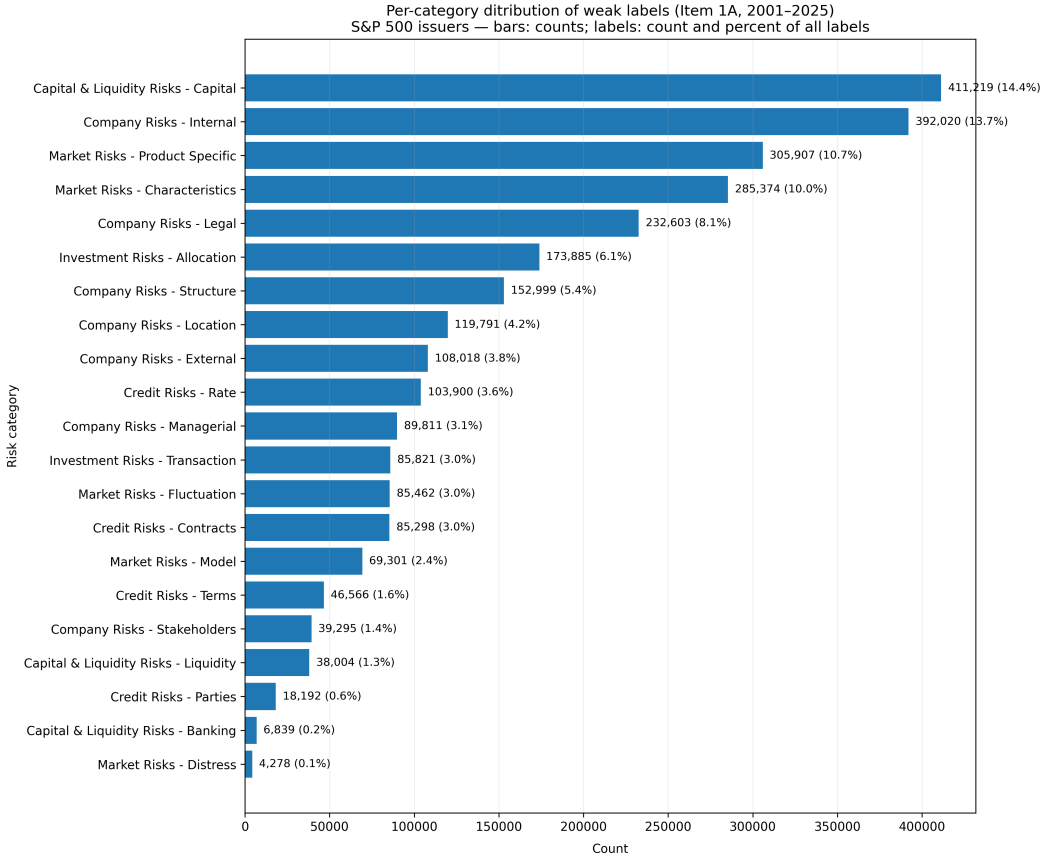

Figure 2: Per-category distribution of weakly-labeled risk sentences (S&P 500, 2001–2025). Bars show counts; labels indicate count and share of all labels.

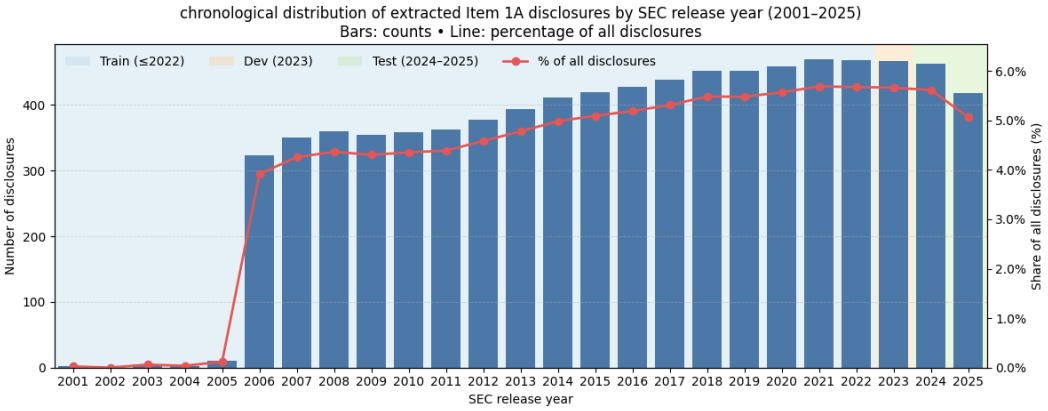

Figure 3: Yearly distribution of extracted Item 1A *disclosures* by SEC release year (2001–2025). Bars show counts; the line shows the percentage of all disclosures. Shaded bands indicate Train (≤2022), Dev (2023), and Test (2024–2025).

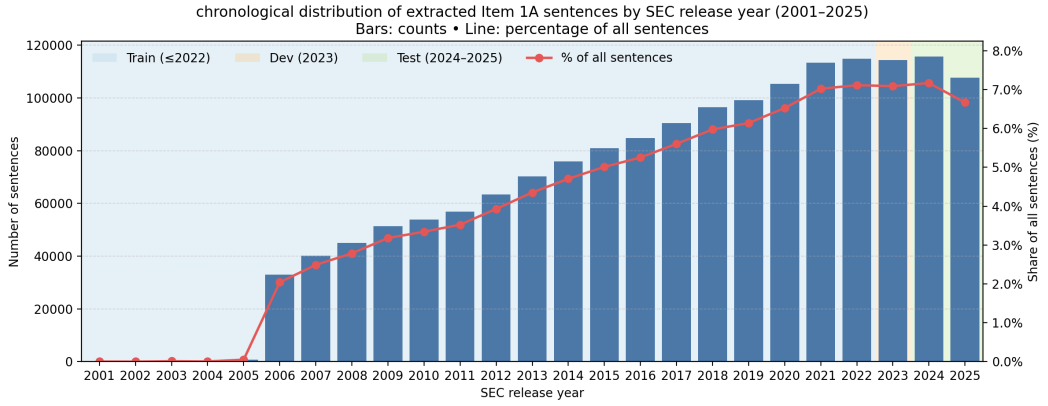

Figure 4: Yearly distribution of extracted Item 1A *sentences* by SEC release year (2001–2025). Bars show sentence counts; the line shows the percentage of all sentences. Shaded bands indicate Train (≤2022), Dev (2023), and Test (2024–2025).