From Lengthy Narrative to Structured Data: Instruction Fine-Tuning Open-Weight LLMs for Information Extraction from Corporate Disclosures

Anonymous ACL submission

Abstract

Texts from regulatory filings contain unstructured information valuable for downstream research, yet extracting it at scale remains costly and unreliable: manual coding is expensive, while prompt-based LLM workflows are sensitive to prompt wording and long context and could degrade sharply when domain-specific reasoning is necessary. To extract structured records from long, heterogeneous corporate disclosures, this paper develops a pipeline organized in two layers: an outer layer that handles term-based retrieval, chunking, extraction, post-processing validation, and cross-chunk aggregation; and an inner layer that fine-tunes the extraction LLM through an iterative error-driven cycle. Using compensation consultant extraction from proxy statements as a testbed—a task requiring long-context comprehension, entity disambiguation, role classification, and temporal reasoning—instruction fine-tuning on 2,001 human-labeled and augmented examples yields F1 of .96, compared with .64–.70 for the open-weight base models and .81–.88 for commercial APIs (GPT-5-nano/mini and GPT-5.5). A Gemma 3 12B variant fine-tuned on the same data nearly matches the 27B one, substituting for base-model scale on a model that fits a single consumer GPU. In addition, models were fine-tuned with both detailed and minimal instructions as separate adapters. Fine-tuned models achieve stable performance regardless of instruction variants (within 2.7 pp), whereas most commercial models benefit from the detailed instruction. This paper releases code for document processing as well as fine-tuned adapters to support adaptation to other long-document extraction tasks.[1]

1 Introduction

Large-scale empirical research increasingly depends on structured data extracted from unstructured archival documents—regulatory filings, court records, and corporate disclosures (Gentzkow et al., 2019; Grimmer et al., 2022). These documents contain fine-grained information unavailable in numerical databases, but transforming their lengthy, heterogeneous narratives into reliable structured records remains a persistent bottleneck. Manual coding is costly and difficult to scale; dictionary-based methods capture only surface patterns.

Large language models (LLMs) offer a promising alternative. Recent work demonstrates that LLMs can match or exceed crowd-worker quality on text annotation tasks at a fraction of the cost (Gilardi et al., 2023; Törnberg, 2023), and general information extraction (IE) alignment approaches (Qi et al., 2024; Sainz et al., 2024; Wang et al., 2023) have advanced extraction capabilities of LLMs on standard benchmarks. However, these benchmarks predominantly feature well-structured, short-context tasks with clear entity boundaries. When applied to long archival documents from specific domains, both prompt engineering and general IE alignment tend to degrade substantially.

This paper addresses this gap by developing a pipeline for extracting structured information from long, heterogeneous documents using instruction fine-tuning (IFT) of open-weight LLMs. The pipeline is organized in two layers: an outer layer handling term-based retrieval, chunking, extraction, post-processing validation, and cross-chunk aggregation at the document level; and an inner layer that fine-tunes the extraction LLM through an iterative error-driven cycle. The pipeline is demonstrated by extracting information on compensation consultants from proxy statements, a task that exhibits the following four challenges common to archival IE:

- **Long-context comprehension**: synthesizing information dispersed across multiple sections of documents spanning as many as hundreds of thousands of tokens.
- **Entity disambiguation**: resolving variable

[1] https://anonymous.4open.science/r/2026submission/

naming conventions and distinguishing target entities from structurally similar distractors.

- **Role classification**: assigning each identified entity to a domain-specific functional category.

- **Temporal reasoning**: distinguishing current relationships from historical references using verb tense, date markers, and contextual cues.

The experiments reveal four findings. First, IFT on 2,001 human-labeled and augmented examples lifts the open-weight Gemma 3 27B model to F1 of .96—a +25.7 percentage point (pp) improvement over its base—and exceeds the frontier GPT-5.5 by +7.7 pp, demonstrating that fine-tuning on a modest amount of quality data significantly improves performance even against the latest frontier model. Second, a smaller Gemma 3 12B fine-tuned model nearly matches the 27B variant (.957 vs. .961 F1) despite their unaligned bases differing by 6.9 pp—suggesting that targeted alignment on a moderately sized open model substitutes for scaling base parameters. Third, scaling at the closed-source frontier shows sharply diminishing returns: moving from GPT-5-mini to GPT-5.5 buys only +0.8 pp on this task (.876 → .884), and even the largest commercial model trails the 12B fine-tuned model. Fourth, comparing two instruction conditions—detailed and minimal—reveals that instruction-length sensitivity declines along two independent axes: base-model scale and fine-tuning. Scaling the open-weight Gemma 3 base from 12B to 27B reduces the minimal-vs-detailed swing from 24.2 pp to 4.3 pp; fine-tuning each model reduces this swing further, to 2.7 pp at 12B and essentially zero at 27B (0.2 pp in the opposite direction).

Beyond these empirical contributions, this paper articulates which pipeline components generalize to other long-document IE tasks: the two-layer architecture, the iterative error-driven fine-tuning cycle, the targeted retrieval-and-chunking strategy, and the output-grounding checks. Other researchers with domain knowledge can replicate the pipeline by supplying task-specific components such as keyword lists, role taxonomies, name-normalization rules, exclusion criteria, and a few hundred human-labeled seed examples. This paper releases code for document processing as well as fine-tuned adapters to support adaptation to other long-document extraction tasks.

2 Related Work

2.1 IE Alignment for LLMs

Recent work aligns LLMs for IE through instruction tuning and multi-task learning. ADELIE (Qi et al., 2024) introduces a multi-stage alignment pipeline; GoLLIE (Sainz et al., 2024) formulates IE as code generation; InstructUIE (Wang et al., 2023) proposes unified instruction tuning across 32 IE datasets; and UniversalNER (Zhou et al., 2024) uses targeted distillation to train models that outperform ChatGPT on out-of-domain NER. These approaches advance general IE capabilities and increasingly adopt generative formulations, where models produce structured outputs directly rather than simply classifying spans (Xu et al., 2024; Zhang et al., 2025). However, these studies focus on general IE benchmarks with well-structured brief texts and relatively simple schemas, leaving open the question of how such methods perform on lengthy, domain-specific texts.

Processing long contexts remains unreliable even for models that can handle large context windows. Liu et al. (2024) identify the "lost-in-the-middle" phenomenon, and Hsieh et al. (2024) show that despite near-perfect accuracy on simple needle-in-a-haystack retrieval, almost all models exhibit large performance drops on more complex tasks as context length increases.

2.2 LLMs as Annotators

LLMs increasingly serve as text annotators (Ziems et al., 2024). Gilardi et al. (2023) and Törnberg (2023) show that ChatGPT outperforms crowd workers on annotation tasks; Heseltine and von Hohenberg (2024) show that GPT-4 can substitute for expert annotators on political text classification, with hybrid validation further raising accuracy; and Goel et al. (2023) show LLMs accelerate annotation for medical IE. Törnberg (2024) synthesizes these findings into guidelines for robust annotation practices, while Alizadeh et al. (2024) show that fine-tuned open-source models offer a competitive alternative to proprietary GPT on text annotation, though still trailing on average.

However, methodological cautions temper this optimism. Egami et al. (2023) show that even high-accuracy LLM predictions introduce substantial bias in downstream analyses. Baumann et al. (2025) replicate 37 annotation tasks across 18 models, revealing that implementation choices can manipulate statistical conclusions. Mu et al. (2024)

find that zero-shot LLMs fail to match fine-tuned BERT on social science classification, with prompt variations causing sizable F1 swings.

2.3 Domain-specific NLP in Finance and Corporate Governance

Domain adaptation is an active direction in financial NLP (Lee et al., 2024). BloombergGPT (Wu et al., 2023) pre-trains a 50B-parameter model on financial text from scratch; FinGPT (Yang et al., 2023) shows that LoRA enables cost-effective adaptation of general LLMs to financial tasks. Closer to this paper's setting, TAT-LLM (Zhu et al., 2024) fine-tunes LLaMA 2 for question answering over financial reports, and RealKIE (Townsend et al., 2024) provides enterprise KIE datasets including SEC filings, observing that existing IE benchmarks "lack realism." Wang and Wang (2024) further argue that LLMs can democratize access to otherwise costly financial datasets, a goal this pipeline shares.

2.4 Bridging the Gaps

This paper synthesizes the three threads above into a single pipeline. To address long-context unreliability, it employs targeted retrieval and chunking that processes relevant spans across multiple passes rather than relying on single-pass extraction over full documents. To tackle prompt brittleness in LLM annotation, it also adopts fine-tuning—an approach that matches or exceeds prompted LLM performance with as few as 1,000 examples (Licht et al., 2025). Human-labeled examples seed an iterative error-driven cycle that expands training data by validating and correcting model outputs, applying the data augmentation logic of synthetic generation (Nayak et al., 2024) while maintaining the quality control of human annotation. The result is a replicable instruction-fine-tuning pipeline for a domain-specific IE task that integrates financial domain knowledge into the language model.

3 Application Setting

3.1 Compensation Committees and the Consulting Industry

In U.S.-listed firms, the compensation committee oversees one of the board's most scrutinized responsibilities: designing, recommending, and approving executive pay. As CEO compensation has grown in both level and complexity, committees must exercise independent judgment while benchmarking against peers and market practices. Gathering comprehensive, comparable evidence on pay design is costly, so many committees retain external compensation consultants (Cadman et al., 2010; Murphy and Sandino, 2020). These consultants advise on compensation architecture and market positioning, provide proprietary surveys, and peer-group benchmarking that go beyond public filings to assist boards in negotiating compensation contracts with executives.

Academic work highlights competing views on compensation consultants. On one side, consultants supply scarce expertise and data; on the other, their conflicts of interest may compromise independence, potentially contributing to increases in CEO pay (Bebchuk and Fried, 2003). Empirically, the evidence is mixed: Some studies associate consultant involvement with higher pay, while others find limited effects after statistical controls (Cadman et al., 2010; Cho et al., 2020; Chu et al., 2018; Murphy and Sandino, 2020).

3.2 Regulatory Responses

The disclosure regime relevant to this paper originated in 2006. That year, the Securities and Exchange Commission (SEC) overhauled executive-compensation disclosure by adding the Compensation Discussion and Analysis (CD&A) section to the proxy statement and requiring narrative discussion of committee processes, such as the use of compensation consultants.[2] In 2010, the SEC's *Proxy Disclosure Enhancements* expanded disclosure to cover consultant fees and potential conflicts when consultants also provided other services above $120,000.[3] This evolving disclosure regime provides the textual basis for the extraction task in this paper.

4 Task and Data

4.1 Task Formulation

This paper formulates compensation consultant extraction as a generative IE task. The fine-tuned language model in the pipeline aims to list the names of compensation consultants currently retained by the firm (RET) and those that provide only benchmarking data or surveys (SURV) from company proxy statements. Extracting information on compensation consultants from proxy statements poses

[2] SEC, *Executive Compensation and Related Person Disclosure*, Release No. 33-8732A (Sept. 8, 2006)

[3] Release No. 33-9089

four challenges that recur across many archival-data IE tasks in general:

Long-context comprehension. Consultant references are dispersed across lengthy proxy statements—sometimes appearing in an introductory list of advisors, later in the pay-philosophy narrative, and again in a peer-benchmarking discussion. Disclosure conventions vary across issuers and time, and editing artifacts (stale text from prior filings, shifting section structures) persist. Building an accurate record requires synthesizing dispersed mentions rather than coding any single passage.

Entity disambiguation. The same consultant appears under multiple names within and across documents (e.g., "Frederic W. Cook & Co.," "FW Cook," "FWC," "Cook & Co."), sometimes with misspellings. Extraction must resolve variants to canonical entities while rejecting structurally similar distractors—auditors and legal advisors that appear in adjacent text. Some firms (e.g., Deloitte, PwC) have historically provided both audit and compensation consulting services, so a name match alone cannot determine whether an entity plays a compensation-consulting role and should be determined based on the reading.

Role classification. Engagement patterns are heterogeneous: a committee may engage a single consultant, retain multiple consultants with distinct scopes, purchase only survey data without an advisory engagement, or report no consultant at all. The RET/SURV distinction carries governance implications regarding independence and conflicts of interest, but disclosures often describe these relationships in adjacent text, requiring careful inference about which consultant plays which role.

Temporal reasoning. Proxy statements reference prior-year consultants for context or simply append new sentences to the previous year's filing. Distinguishing current from historical engagements thus requires parsing verb tense, date markers (e.g., "in FY 2023"), and phrases such as "formerly" or "previously." Without temporal grounding, extraction produces duplicate or stale records.

These challenges explain why prior work relied on small hand-collected samples or proprietary databases (Cadman et al., 2010; Cho et al., 2020; Murphy and Sandino, 2020). These challenges are also not unique to consultant disclosures—long-context comprehension, entity disambiguation, role classification, and temporal reasoning recur across corporate-disclosure and other archival IE tasks, motivating a generalizable pipeline.

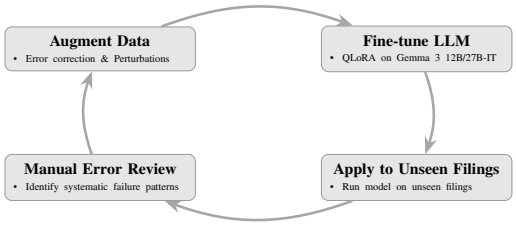

Figure 1: Iterative Error-driven Cycle for Training Data Construction and Fine-tuning. Human-labeled seed examples initialize fine-tuning. Errors on new documents are manually corrected and augmented with controlled perturbations, then added to the training set for the next iteration.

4.2 Data Construction and Fine-tuning

This paper builds the training set and fine-tunes the LLM jointly through an iterative error-driven cycle—the *inner layer* of the pipeline—rather than assembling either in a single pass (Figure 1). The resulting model is what the outer layer applies to each retrieved chunk. Starting from 400 human-labeled seed examples drawn from proxy statements spanning 2007–2025, each iteration proceeds through four stages: (1) fine-tuning on the current training set; (2) application to unseen filings; (3) manual error review to identify systematic failure patterns; and (4) augmentation via error correction and controlled perturbations (entity substitution, sentence reordering, distractor injection). The loop terminates once performance stabilizes on a held-out monitoring set, at which point the training set comprises 2,001 examples, split 80/20 for training and validation. A separately labeled set of 143 proxy statements from 84 companies, containing 316 consultant labels (159 RET, 157 SURV), is held out as the test set and never seen during the cycle. An example training instance is shown in Appendix A.

The fine-tuning configuration is held fixed across iterations. Fine-tuning targets Gemma 3 at two parameter scales—12B-Instruct and 27B-Instruct (Gemma Team, 2025)—using QLoRA (Dettmers et al., 2023) with 4-bit NF4 quantization. LoRA adapters (Hu et al., 2021) are applied to attention projections (q, k, v, o) and MLP layers (gate, up, down) with rank $r = 8$, $\alpha = 16$, and dropout 0.05. Training uses a two-turn conversational format with the loss computed only on the model's output turn. The maximum sequence length is set to 5,120 tokens—up to 4,000 for a retrieved text chunk plus the instruction and a small buffer—

matching the output of the retrieval-and-chunking stage; this length applies to both training and inference. The checkpoint with the lowest validation loss is selected for evaluation. Fine-tuning runs on a single NVIDIA DGX Spark workstation (128 GB unified memory).

5 Pipeline Architecture

The outer layer of the pipeline (Figure 2) consists of four stages—retrieval, extraction, post-processing, and aggregation—and is designed to transfer to other long-document IE tasks. Adapting it to a new domain requires replacing the task-specific components (keywords, chunking rules, post-processing rules) while preserving the architecture.

Targeted Retrieval and Chunking. Rather than passing entire documents to the model, candidate spans are identified via keyword matching, extended with 250 words of surrounding context on each side, merged where overlapping, and segmented into chunks of up to 4,000 tokens. This addresses both the "lost-in-the-middle" degradation (Liu et al., 2024) and the memory limits of consumer-grade GPUs.

Extraction. Each chunk is passed to the fine-tuned model with a standardized task prompt and low-temperature decoding for reproducibility. The paper evaluates two prompt variants—detailed and minimal—compared in Section 7.

Post-processing. Grounding checks verify that extracted names appear in the source chunk, preventing hallucination (Ji et al., 2023). Name normalization maps variants to canonical forms (e.g., "FW Cook," "FWC," "Frederic W. Cook & Co." → the longest variant found in each document).

Cross-chunk Aggregation. Chunk-level outputs are combined into document-level records: name variants are resolved; role conflicts are reconciled (RET supersedes SURV for dual-role consultants); and placeholder labels ("UNNAMED" and "NO_CONSULTANTS") are replaced with named extractions from other chunks of the same proxy statement.

6 Experiments

6.1 Experimental Setup

Models. The following configurations are compared:

Table 1: Main Results on Compensation Consultant Extraction. Best result in **bold**.

Model	Type	P	R	F1
Gemma 3 12B-IT	zero-shot	.614	.658	.635
Gemma 3 27B-IT	zero-shot	.656	.759	.704
+ General IE (NER)	fine-tuned	.697	.832	.758
+ General IE (Bal.)	fine-tuned	.841	.864	.852
+ Domain-Specific 12B	fine-tuned	.958	.956	.957
+ Domain-Specific 27B	fine-tuned	.957	.965	**.961**†
GPT-5-nano	zero-shot	.800	.810	.805†
GPT-5-mini	zero-shot	.882	.870	.876
GPT-5.5 (frontier)	zero-shot	.888	.880	.884

†Evaluated with minimal instructions, the better-performing setting for these models; all other models are shown under detailed instructions.

- **Base (zero shot)**: Gemma 3 12B-IT and 27B-IT
- **General IE (NER)**: Gemma 3 27B fine-tuned on 2,001 NER examples from IEInstruct (Qi et al., 2024)
- **General IE (Balanced)**: Gemma 3 27B fine-tuned on 2,001 IEInstruct examples (800 NER, 800 relation extraction, 401 event extraction)
- **Domain-Specific**: Gemma 3 12B and 27B fine-tuned on the 2,001 in-domain examples, with parallel detailed- and minimal-instruction adapters
- **Commercial APIs (zero shot)**: GPT-5-nano, GPT-5-mini, and GPT-5.5

Each Domain-Specific adapter is evaluated with the instruction style it was trained on; cross-style results appear in Section 7.

Evaluation. Evaluation uses 143 held-out proxy statements from 84 companies, containing 316 consultant labels (159 RET, 157 SURV). Precision, recall, and F1 are reported at the instance level: each consultant must be correctly identified and categorized.

Hardware. Inference experiments were run on a desktop workstation with two 24 GB GPUs.

6.2 Main Results

Table 1 presents the main results. Each model is reported under its better-performing instruction condition (see Section 7); the dagger marks models where this is minimal rather than detailed. Four findings emerge:

Domain-specific alignment yields substantial gains. The 27B fine-tuned model achieves 96.1% F1, a +25.7 percentage point improvement over the 27B base and +10.9 pp over the best general IE baseline. This demonstrates that domain-specific

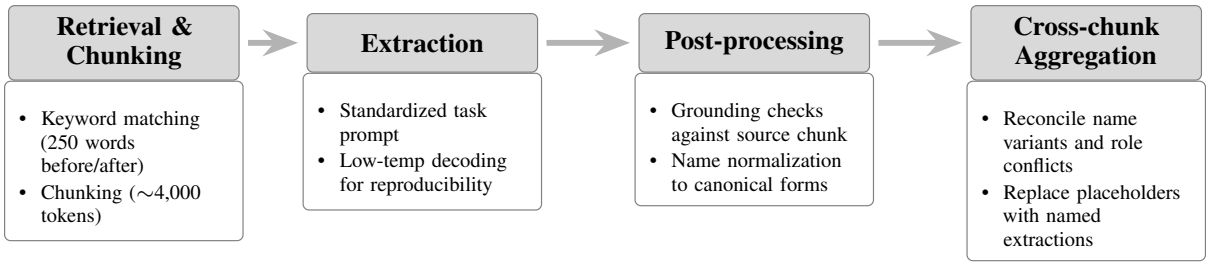

Figure 2: Multi-stage Extraction Pipeline. The pipeline processes each document through four stages; only the extraction model varies across experimental conditions.

training data is necessary for IE tasks requiring nuanced reasoning coupled with domain expertise.

The 12B fine-tuned model matches the 27B variant. Fine-tuning Gemma 3 12B on the same 2,001 examples reaches 95.7% F1—within 0.4 pp of the 27B fine-tuned model—despite their unaligned bases differing by 6.9 pp (12B base .635 vs. 27B base .704). Targeted alignment on a moderately sized open-weight model substitutes for base-model scale, meaning researchers can deploy on more accessible hardware without sacrificing accuracy.

General IE alignment helps but plateaus. Fine-tuning on general IE data improves on base zero-shot performance (NER: +5.4 pp; Balanced: +14.8 pp), but remains well below domain-specific performance. Task diversity matters—the balanced mixture outperforms NER-only by 9.4 pp—but general IE training teaches *how* to extract, not *what* distinctions matter in a specific domain.

Domain-specific fine-tuning of small open models outperforms frontier commercial APIs. The fine-tuned open-weight 27B exceeds the frontier GPT-5.5 by 7.7 pp, with even the smaller 12B variant beating GPT-5-mini (Table 1). Scaling along the closed-source ladder from GPT-5-mini to GPT-5.5 buys only +0.8 pp (.876 → .884)—near-zero marginal benefit at the frontier, whereas targeted fine-tuning on 2,001 examples produces a much larger gain.

6.3 Category-Level Analysis

Table 2 disaggregates performance by consultant category. All models perform better on SURV than RET, likely because survey-only relationships involve clearer linguistic markers. By contrast, RET classification requires subtle reasoning about advisory scope and committee engagement—precisely the domain knowledge encoded in the training data.

Table 2: Performance by Consultant Category

Model	RET F1	SURV F1	Δ
Base 27B (zero-shot)	.700	.708	+.008
General IE (Balanced)	.831	.875	+.044
Domain-Specific 12B	.937	.978	+.041
Domain-Specific 27B	.957	.965	+.008
GPT-5-nano	.784	.827	+.043
GPT-5-mini	.833	.920	+.087

Domain-specific alignment provides the largest absolute gains on RET (+25.7 pp over the 27B base), the more challenging category. Notably, the RET–SURV gap nearly disappears for the 27B fine-tuned model ($\Delta = +.008$), whereas it widens for commercial APIs (+.087 for GPT-5-mini)—suggesting that targeted alignment specifically internalizes the domain reasoning needed to identify retained advisors.

7 Ablation Studies

7.1 Effect of Instruction Length

This subsection investigates whether performance gains stem from instruction engineering or from alignment on domain-specific reasoning. The two prompt conditions are:

- **Detailed** (∼860 tokens): Full specification of temporal logic, exclusion criteria, role definitions, and edge cases with examples
- **Minimal** (∼265 tokens): Core task definition only

The minimal instruction variant is shown in Appendix B.

Table 3 reveals that instruction-length sensitivity declines along two independent axes—base-model scale and fine-tuning—and that the two can interact.

Table 3: Effect of Instruction Length on F1. Δ = Minimal − Detailed; a negative value means detailed instructions help.

Model	Type	Det.	Min.	Δ
Gemma 3 12B-IT	zero-shot	.635	.393	−.242
Gemma 3 27B-IT	zero-shot	.704	.661	−.043
GPT-5-nano	zero-shot	.608	.805	+.197
GPT-5-mini	zero-shot	.876	.813	−.063
GPT-5.5 (frontier)	zero-shot	.884	.804	−.080
General IE (NER)	fine-tuned	.758	.742	−.016
General IE (Bal.)	fine-tuned	.852	.774	−.078
Domain-Specific 12B	fine-tuned	.957	.930	−.027
Domain-Specific 27B	fine-tuned	.959	.961	+.002

Small and unaligned models benefit most from detailed instructions. The 12B base swings by 24.2 pp between minimal and detailed instructions, while every other model gains far less: the 27B base by 4.3 pp, GPT-5-mini by 6.3 pp, and GPT-5.5 by 8.0 pp.[4] The smaller the unaligned model, the more it depends on detailed prompts as scaffolding.

General IE alignment partially internalizes instructions. Both general IE models still gain from detailed prompts (NER +1.6 pp, Balanced +7.8 pp)—less than the unaligned bases but more than domain-aligned models, placing them between the two on the sensitivity axis.

Fine-tuning largely eliminates instruction-length sensitivity. The 27B fine-tuned model is instruction-length-invariant (0.2 pp gap in the opposite direction). The 12B fine-tuned model retains only a 2.7 pp gap, 89% smaller than the 24.2 pp swing of its unaligned base. Scale and alignment substitute partially for one another: a larger model needs little instruction scaffolding once aligned, while a smaller aligned model still benefits marginally. The interaction is visible across rows: fine-tuning's effect grows with base scale, eliminating the residual gap entirely at 27B but only shrinking it at 12B.

Implication. Detailed instructions improve weaker approaches but cannot match domain-specific fine-tuning. For research applications requiring reproducibility and stability, annotation quality matters more than prompt engineering.

[4]GPT-5-nano reverses the pattern (−19.7 pp under detailed instructions), which likely reflects weaker instruction-following at the smallest commercial scale.

Table 4: Effect of General IE Training Composition

Training Data	F1	Δ vs. Base
Base 27B (no fine-tuning)	.704	—
NER only	.758	+.054
Balanced (NER + EE + RE)	.852	+.148
Domain-Specific	.961	+.257

7.2 Effect of General IE Training Composition

The Balanced vs. NER-only gap (Table 4, 9.4 pp) isolates the marginal contribution of event and relation diversity beyond entity recognition alone—and even the best general IE configuration still falls 10.9 pp short of domain-specific alignment, which teaches the model to resolve the four extraction challenges identified in Section 4.

8 Discussion

8.1 When Is Domain-Specific Fine-Tuning Necessary?

General IE alignment may suffice for tasks with simple formats, clear entity boundaries, and minimal domain knowledge requirements. These results suggest that domain-specific IFT is necessary when extraction tasks exhibit the four challenges introduced in Section 4: (1) long-context comprehension across many pages of disclosure; (2) entity disambiguation among surface-similar candidates; (3) role classification requiring domain-specific reasoning about what each entity does in context; and (4) temporal reasoning to separate current from historical mentions.

8.2 Practical Implications

For NLP practitioners building IE systems for archival documents, the results above offer three actionable insights.

First, **invest in annotation quality over prompt engineering**: zero-shot performance depends heavily on instruction length (4–24 pp swings); fine-tuning removes this dependency.

Second, **targeted alignment substitutes largely for base-model scale**: a 12B fine-tuned model matches the 27B variant (.957 vs. .961 F1) despite a 6.9 pp gap between their unaligned bases.

Third, **error-driven iteration concentrates annotation effort**: growing the training set from 400 seed to 2,001 examples through targeted error correction and perturbation augmentation lets each

added example resolve an observed failure mode rather than pad redundant coverage.

8.3 Generalization

Although demonstrated here on regulatory filings, the methodology should generalize to long-document IE tasks that share the four extraction challenges. Candidate applications include IE from other SEC disclosures in financial studies, clause-level IE in legal NLP, and policy coding from corporate annual reports in management research. Adopting the pipeline requires two ingredients: a few hundred human-labeled examples as a seed for the iterative error-driven cycle, and keywords or heuristics (or a lightweight fine-tuned retriever, for less keyword-friendly domains) that retrieve candidate chunks from the document collection.

Limitations

Single-domain evaluation. Evaluation covers one extraction task from one document type (SEC proxy statements). Although the framework is argued conceptually to generalize, this is not demonstrated empirically on a second domain—the primary limitation of this paper.

Single model family. Fine-tuning experiments cover only the Gemma 3 family at two parameter scales (12B and 27B). Testing fine-tuning on additional architectures (e.g., LLaMA, Mistral) would strengthen claims about architecture-independence.

Test set size. The held-out test set of 143 filings from 84 companies / 316 consultant labels, while drawn from firms and years excluded from training, is modest. Cross-validation across multiple splits would provide stronger statistical evidence.

Context length constraints. The 5,120-token context window, matched across all models, may not capture all relevant information in very long filings. Techniques such as position interpolation (Chen et al., 2023) and YaRN (Peng et al., 2023) enable inference on longer sequences, but their effectiveness for domain-specific IE tasks remains unexplored.

Ethics Statement

This work processes publicly available regulatory filings and does not involve private or sensitive data. The pipeline is designed to extract factual information from long-context corporate disclosures; it does not generate persuasive or deceptive content. Released resources support reproducibility and reduce barriers for researchers with limited computational resources.

Acknowledgments

AI assistants were used for editorial revisions, prose suggestions, and code reasoning support during paper preparation. All substantive research contributions, analyses, and conclusions are the author's own.

[Personal acknowledgments — anonymized for review.]

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Simin Dehghani, Moe Zahedivafa, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2024. Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*.

Leah Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. *arXiv preprint arXiv:2509.08825*.

Lucian Arye Bebchuk and Jesse M. Fried. 2003. Executive compensation as an agency problem. *Journal of Economic Perspectives*, 17(3):71–92.

Brian Cadman, Mary Ellen Carter, and Stephen Hillegeist. 2010. The incentives of compensation consultants and CEO pay. *Journal of Accounting and Economics*, 49(3):263–280.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via position interpolation. *arXiv preprint arXiv:2306.15595*.

Jae-Ho Cho, Inwoo Hwang, Jung-Ho Hyun, and Jae Yong Shin. 2020. Compensation consultant fees and CEO pay. *Journal of Management Accounting Research*, 32(1):51–78.

Jenny Chu, Jonathan Faasse, and P. Raghavendra Rau. 2018. Do compensation consultants enable higher CEO pay? A disclosure rule change as a separating device. *Management Science*, 64(10):4915–4935.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. Using imperfect surrogates

for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Matthew Gentzkow, Bryan T. Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature*, 57(3):535–574.

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.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. LLMs accelerate annotation for medical information extraction. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1).

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A survey of large language models in finance (FinLLMs). *arXiv preprint arXiv:2402.02315*.

Hauke Licht, Rupak Sarkar, Patrick Y. Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. Measuring scalar constructs in social science with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32144–32171, Suzhou, China. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086. ELRA and ICCL.

Kevin J. Murphy and Tatiana Sandino. 2020. Compensation consultants and the level, composition, and complexity of CEO pay. *The Accounting Review*, 95(1):311–341.

Nihal V. Nayak, Yiwei Wang, Jesse Dodge, and Noah A. Smith. 2024. Bonito: Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shao. 2023. YaRN: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Yunjia Qi, Hao Tang, Zixuan Wan, Chen Liang, Zhiyang Shen, Xiangyu Zhao, Chao Zhang, Michael R. Lyu, and Irwin King. 2024. ADELIE: Aligning large language models on information extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7357–7393. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Rodrigo Agerri, Oier Lopez de Lacalle, and Eneko Agirre. 2024. GoLLIE: Annotation guideline improves zero-shot information extraction. In *The Twelfth International Conference on Learning Representations*.

Petter Törnberg. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Petter Törnberg. 2024. Best practices for text annotation with large language models. *Sociologica*.

Benjamin Townsend, Madison May, and Christopher Wells. 2024. RealKIE: Five novel datasets for enterprise key information extraction. *arXiv preprint arXiv:2403.20101*.

Jingwei Wang and Vincent X. Wang. 2024. Leveraging large language models to democratize access to costly financial datasets for academic research. *arXiv preprint arXiv:2412.02065*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Jian Ye, Qi Zhang, Tao Gui, Jihua Li, Jingyi Yang, Xuanjing Huang, and Jiang Bian. 2023. InstructUIE: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Xiaoyu Zhu, Tong Gao, Haoran Chen, Zhiyu He, Dong Wang, Shuhan Zhang, and Peng Zhang. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-source financial large language models. In *FinLLM Workshop @ IJCAI 2023*.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025. A survey of generative information extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhuang, Noah A. Smith, and Hannaneh Hajishirzi. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat-Seng Chua. 2024. TAT-LLM: A specialized language model for discrete reasoning over tabular and textual data. In *ICAIF 2024*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Training Example

Below is a sample training instance formatted as a two-turn conversation. The user turn contains the task instruction and a proxy statement excerpt; the model turn contains the structured extraction output.

User Turn (Input)

```
### Instruction ###
You are a meticulous reader of proxy-statement excerpts.

TASK
Your task is to identify two categories of compensation consultants
mentioned in the text:
  1. RET: Consultants retained/engaged/selected/hired by the committee
     or management to advise on executive compensation (e.g., conduct
     peer group analysis, write up a report).
  2. SURV: Consultants that provided compensation surveys and/or
     benchmarking/market data but were NOT explicitly retained/engaged.
     If there is a reference that a retained entity explicitly provides
     those, list it also under SURV.

INPUT FORMAT
You will receive a collection of excerpts separated by the token
<<<EXCERPT_BREAK>>>. Treat them as one document.

STRICT OUTPUT RULES (CUSTOM FORMAT; NOT JSON):
1. Return two dictionaries, exactly in this order and with these keys:
   {RET: 'Firm A', 'Firm B'}, {SURV: 'Firm C'}
2. When a firm appears under multiple names, output only the LONGEST
   name found in the whole document/excerpts.
3. List the firms in alphabetical order in each dictionary, with every
   firm enclosed in single quotes.
4. If no firms are found for a category, use NO_CONSULTANTS. If a
   firm's name is not disclosed, use 'UNNAMED'.

IMPORTANT TIMING RULES:
- You will see a filing year at the start of each document. If a
  following excerpt/context contains no specific year, assume the
  information is current.
- Fiscal year: Proxy statement may refer to a fiscal year. A fiscal
  year is 12 months but starts later than January 1, depending on
  business needs or regulatory requirements. Thus, Fiscal Year 2022
  could refer to a period in 2023.
- A consultant counts as currently retained if they were actively
  retained at the time of or during the filing year (or the immediately
  prior year), and had not been replaced before the filing date. If
  the proxy explicitly states a consultant was replaced, then only the
  last consultant should be counted even if the predecessor was retained
  in the filing or previous year.

TEMPORAL CONTEXT: This document was filed in 2020. Consultants retained
in 2020 or 2019 are considered 'currently retained'.

TEXT TO ANALYZE:

Company's compensation plans, policies, and programs. The
Compensation Committee's duties and responsibilities include
reviewing and approving on an annual basis corporate goals and
objectives relevant to compensation of the CEO and other members
of the executive team as well as approving all grants of equity-
related compensation. The Compensation Committee, together with
the ESG Committee, reviews the Company's Organization and
Succession Plan at least annually.

The current members of the Compensation Committee are Scott
(Chair), Jerri, Mike, Don, Greg, Bala, and Bryan. Our Board
has determined that each member of the Compensation Committee
meets the independence requirements of the SEC as well as
those of the NYSE. The Compensation Committee met four times
in 2020. The Compensation Committee also held one joint
meeting with the ESG Committee in 2020.
```

The Compensation Committee has primary responsibility for administering our Omnibus Incentive Compensation Plan and in that role is responsible for approving equity grants to employees.

The Compensation Committee also oversees the organization plan and senior officer succession planning in consultation with the ESG Committee.

Under its charter, the Compensation Committee may, in its sole discretion, retain or obtain advice of a compensation consultant, independent legal counsel, or other adviser. The Compensation Committee is directly responsible for the appointment, compensation and oversight of any such consultant, counsel or adviser, and the Company shall provide appropriate funding for payment of reasonable compensation to any such consultant, counsel or adviser, as determined by the Compensation Committee. In selecting a consultant, counsel, or adviser, the Compensation Committee evaluates its independence.

During 2020, the Compensation Committee retained Meridian Compensation Partners, LLC ("Meridian") as its consultant to advise it on executive compensation matters for which Meridian received customary fees. Meridian's work included providing guidance regarding the treatment of equity compensation awards and related matters. After considering factors used by the Compensation Committee to evaluate independence, the Compensation Committee determined that Meridian is an independent compensation consultant in accordance with applicable SEC and NYSE rules.

Meridian participates in Compensation Committee meetings as requested by the Chair of the Compensation Committee and communicates directly with the Chair of the Compensation Committee outside of meetings. Meridian has provided the following services to the Compensation Committee:

<<<EXCERPT_BREAK>>>

performance goals, reviewing succession plans relating to the CEO and our other executive officers, the adoption, amendment and termination of cash and equity-based incentive compensation plans, approving any employment agreements, severance agreements or change in control agreements with executive officers, and the level and form of non-employee director compensation and benefits. Nominating and Governance Committee Overview The Nominating and Governance Committee consists of 5 directors, all of whom are independent under BZX and Nasdaq listing rules. The Nominating and Governance Committee's responsibilities include making recommendations to the Board on: persons for election as director, a director to serve as Chairman of the Board and an independent director to serve as Lead Director, any stockholder proposals and nominations for director, the appropriate structure, operations and composition of the Board and its Committees, the Board and Committee annual self-evaluation process, and the contents of the Corporate Governance Guidelines, Code of Business Conduct and Ethics and other corporate governance policies and programs.

<<<EXCERPT_BREAK>>>

ground, whether the individual has any special experience in a relevant area, personal accomplishments and cultural experiences. In addition, the Committee may consider such other factors it deems appropriate when conducting its assessment of director candidates. Diversity While we do not currently have a formal diversity policy, our Corporate Governance Guidelines provide that the Nominating and Governance Committee will seek to

```
recommend to the Board candidates for director with a diverse
range of experiences, qualifications and skills in order to
provide varied insights and competent guidance regarding our
operations, with a goal of having a Board that reflects diverse
backgrounds, experience and viewpoints.

<<<EXCERPT_BREAK>>>

ected as a director once he or she reaches age 73. Any director
who turns 73 while serving as a director may continue to serve
for the remainder of their current term. Cboe Global Markets
2020 Proxy Statement 19 Table of Contents Annual Board and
Committee Self-Evaluations The Board believes that a robust
annual evaluation process is a critical part of its governance
practices.

<<<EXCERPT_BREAK>>>

u may ABSTAIN. Each nominee must receive the affirmative vote
of a majority of the votes cast with respect to his or her
election in order to be elected. Each nominee has tendered his
or her resignation, contingent on failing to receive a majority
of the votes cast in this election and acceptance by the Board.

<<<EXCERPT_BREAK>>>

The Audit Committee reviewed the independence of Hostel, LLP.
and concluded that it is an independent legal counsel.
```

Model Turn (Output)

```
{RET: 'Meridian Compensation Partners, LLC'}, {SURV: 'NO_CONSULTANTS'}
```

Annotation Rationale. Meridian is classified as RET because the committee formally "retained" them to "advise on executive compensation matters." No firms are mentioned as providing survey or benchmarking data without being retained, so SURV returns NO_CONSULTANTS. The excerpt also mentions "independent legal counsel" and "other adviser" generically, but these are correctly excluded as they refer to potential future engagements, not compensation consultants.

B Minimal Instruction Prompt

Below is the minimal instruction variant used in ablation studies (Section 7). This version omits temporal logic rules, entity disambiguation criteria, and edge case handling.

```
### Instruction ###
You are a meticulous reader of proxy-statement excerpts.

You will receive a collection of excerpts separated by the token
<<<EXCERPT_BREAK>>>. The excerpts may come from one or several proxy
statements of different companies.

Your task is to identify two categories of compensation-consulting
firms mentioned in the text:
  1. Retained compensation consultants - firms that were formally
     engaged to provide advice and/or data on executive compensation.
  2. Survey-data providers - firms that only sold compensation survey
     or benchmarking data and were not formally retained.

STRICT OUTPUT RULES:
  1. List the firms in alphabetical order, each enclosed in single quotes.
  2. Return two dictionaries, exactly in this order and with these keys:
     {RET: 'Name1', 'Name2', ...}, {SURV: 'Name1', 'Name2', ...}
  3. If no firms are found for a category, use NO_CONSULTANTS as the
     sole entry for that category.

### Proxy Statement Excerpt ###
TEMPORAL CONTEXT: This document was filed in [YEAR], discussing fiscal
year [YEAR-1].
```

```
TEXT TO ANALYZE:
[Document text inserted here]
```