

DECODEM: Data Extraction from Corporate Organizational Documents via Enhanced Methods

Jens Frankenreiter

Washington University in St. Louis School of Law

Draft. Comments welcome.

Abstract

Much empirical legal research depends on translating unstructured text into structured variables. In corporate governance research as elsewhere, this translation has traditionally relied on human coding of documents such as charters and bylaws, a process that is costly, difficult to scale, and often opaque. This paper introduces DECODEM, a set of benchmark datasets for evaluating the automated extraction of corporate governance variables from organizational documents. The benchmarks pair randomly sampled corporate charters and bylaws with high-quality human annotations covering a range of governance provisions commonly studied in empirical work.

Using these datasets, the paper evaluates several large-language-model extraction pipelines that vary in prompt design, task decomposition, and document handling. The underlying task consists of a set of document-level binary classification problems, one for each governance variable. The results show that automated extraction is feasible at a high level of accuracy for many provisions, with median performance near the upper bound across approaches. At the same time, performance varies systematically across variables, with a small number of provisions accounting for most of the remaining errors. More elaborate prompting strategies and cascading pipelines do not consistently improve performance for frontier models, but substantially narrow the gap between frontier and efficiency-oriented models in some settings, suggesting that pipeline design can partly substitute for model capability.

By providing a standardized benchmark and a systematic evaluation of extraction methods, the paper demonstrates that current frontier models can extract legally meaningful information from complex corporate documents with high accuracy and suggests an important future role for automated feature extraction in constructing corporate governance datasets.

1 Introduction

Much empirical legal research depends on translating unstructured text into structured variables. A prominent example is research in law and finance, where decades of work have examined the real-world implications of corporate governance arrangements (e.g., [4, 17]). Traditionally, this translation has been performed by human coders, who read and code legal documents such as corporate charters and bylaws[20]. While feasible, this approach imposes important constraints on research, limiting the scale and scope of empirical analysis and raising the question whether this step can be automated.

Legal texts, including corporate charters and bylaws, present distinctive challenges for natural language processing. They are often lengthy, structurally complex, and written in specialized language that varies across documents [3, 31]. Moreover, the meaning of key provisions frequently depends on context, and even small differences in phrasing can have significant legal implications. Traditional NLP methods, which tend to perform well on more standardized text, have struggled with this combination of length, variability, and context dependence. The advent of large language models has made large-scale extraction from such documents a plausible prospect. Yet, to date, there is little systematic evidence on whether automated methods can reliably recover legally meaningful variables from corporate charters and bylaws at scale.

This paper introduces two benchmark datasets (the DECODEM benchmarks) designed to evaluate the automated extraction of corporate governance variables from charters and bylaws. The benchmarks are constructed from a corpus of real-world corporate governance documents and focus on provisions governing shareholder meetings, director elections, board structure, indemnification, exculpation, and takeover-related arrangements, which are the types of provisions that have played a central role in empirical research. For each provision, the datasets provide hand-coded labels that capture whether the relevant governance arrangement is present.

Building on these datasets, the paper evaluates a range of extraction pipelines that vary along several dimensions, including the complexity of the prompts used to query the model and the extent to which the model is provided with the full document or with selected excerpts identified through a pre-processing step. This design makes it possible to assess not only overall performance, but also the tradeoffs between different approaches to prompt engineering and document handling.

The results show that large language models can recover many governance variables with high accuracy. At the same time, performance varies substantially across variables. A small number of arguably more complex provisions—such as shareholder meeting rights and removal and vacancy-filling authority—account for most of the observed errors. Importantly, more complex prompting and cascading extraction pipelines do not consistently improve performance for frontier models, but they substantially narrow the gap between frontier and efficiency-oriented models for the bylaws benchmark. This suggests that pipeline design can partly substitute for model capability. At the same time, remaining errors in the strongest models appear to stem less from extraction architecture than from the interpretive demands of particular governance variables.

Overall, the paper makes three distinct contributions. First, it introduces DECODEM, a pair of benchmark datasets for evaluating long-document legal information extraction from corporate charters and bylaws. Second, it provides a systematic comparison of extraction architectures and frontier language models on document-level governance-variable classification tasks. Third, it analyzes where extraction fails, showing that remaining errors are concentrated in particular governance provisions and observations rather than distributed evenly across the benchmark. A companion paper [13] uses the same benchmarks to address a complementary set of questions, comparing automated extraction against realistic human coding workflows and evaluating ensemble coding and LLM-assisted audits of human-coded data.

2 Text-to-Variable Translation in Empirical Legal Research

A central step in many empirical legal research projects is transforming unstructured text into structured variables for statistical analysis[16, 20]. Researchers draw on textual sources—contracts, statutes, regulatory filings, court opinions, or corporate disclosures—to study institutional arrangements and their economic or social consequences, but the relevant information is rarely in a directly analyzable format.

It must first be encoded into structured data, typically by identifying whether particular concepts appear in a document and mapping them to categorical or binary variables under predefined coding rules.

Traditionally this encoding is done by hand: trained coders read documents and label the presence or absence of specific features. Manual coding is time-consuming and costly[20], so datasets tend to cover limited samples, and projects with uncertain payoffs may not be attempted at all. Coding decisions are also interpretive, and even well-designed protocols can produce inconsistencies across coders or over time. Standardized datasets can amortize these costs across many users, but at the price of a fixed menu of variables and limited transparency into how they were coded.

Legal and regulatory documents are often long, structurally complex, and written in specialized language, so relevant information is often spread across clauses or sections and sensitive to small differences in phrasing[5, 39]. This makes rule- and keyword-based extraction challenging. Earlier NLP methods worked when targets could be identified from vocabulary[30, 33], but degraded when extraction required contextual or legally informed interpretation.

Such methods also depend on large labeled training sets, which are usually unavailable here. One response in machine learning has been to reduce reliance on manual annotation through weak or distant supervision. These approaches substitute programmatic or noisy labels for gold-standard annotations, trading label fidelity for scale[29, 32].

Modern large language models offer a different route: they capture contextual relationships, perform tasks requiring semantic understanding, and in many settings need little or no task-specific labeled data[8]. In principle, this allows structured information to be extracted directly from complex documents, automating at least some text-to-structure translation.

Realizing this potential, however, requires benchmarks that pair primary documents with high-quality human annotations. Such benchmarks measure extraction accuracy, enable comparison across approaches, track progress over time, and surface the textual structures that remain hard to interpret automatically. Constructing them is therefore an important step toward automating the extraction of legal information.

3 Empirical Corporate Governance Research and the Data Bottleneck

This paper studies this benchmark problem in the context of empirical corporate governance research, a natural

setting for translating legal text into structured data. A large body of work in law and finance relies on variables capturing the governance arrangements of public corporations—director elections, shareholder rights, takeover defenses, and litigation-related provisions[4, 17]. Much of this information lives in corporate charters and bylaws, which define key elements of a firm’s governance framework but exist almost exclusively as text and must be translated into structured variables before statistical analysis.

This translation has traditionally relied on hand collection[25] or standardized commercial datasets[4, 17]. Both have enabled influential work, but both have limits. Manual coding is costly and hard to scale, especially for features requiring legally informed interpretation or for large longitudinal datasets. Commercial datasets lower costs for end users but typically cover only a subset of variables and offer limited visibility into how those variables were defined and coded. Prior work also questions the accuracy of some widely used governance datasets, suggesting data quality should be a central methodological concern rather than a background assumption[14].

These constraints limit not just how much data is available but the research agenda itself. With only a small set of governance variables in structured form, researchers gravitate toward questions those variables can answer, while many potentially important arrangements remain hard to study at scale because the relevant information has never been systematically extracted. Empirical corporate governance thus remains tied to the scope and quality of legacy data collection.

Automated extraction from charters and bylaws offers a way past these constraints. Reliable large-scale identification and classification of legally relevant provisions would let researchers replicate parts of existing datasets and study arrangements absent from standard sources, broadening the empirical agenda and reducing dependence on opaque coding processes. Recent work shows the promise: computational analysis of large corpora of corporate documents can surface governance practices that are otherwise hard to observe systematically[12, 15].

Charters and bylaws are also a demanding and practically important test case for modern NLP. They are legal in nature, often long, and vary widely across firms and over time, so extraction depends on contextual and legally informed interpretation rather than surface lexical patterns. Benchmarks built from them therefore serve two purposes: they support empirical corporate governance research and provide a broader evaluation setting for legal text-to-structure translation in long, heterogeneous documents.

4 Related Work

This paper contributes to several strands of literature. First, a notable literature in natural language processing has developed benchmark datasets for a wide range of tasks, including legal tasks (e.g., [1, 9, 19, 21]). However, these benchmarks typically focus on tasks designed to assess a model’s ability to perform forms of legal reasoning, classification, or retrieval, rather than on the extraction of structured variables from lengthy organizational documents. By contrast, the extraction problems that arise in empirical legal research often require the processing of long, structurally complex texts and the recovery of legally relevant information in a form suitable for quantitative analysis.

In terms of task design, the closest neighbor to DECODEM is Marotta-Wurgler and Stein’s privacy-policy corpus, which develops a hand-coded dataset and coding toolkit for translating lengthy legal documents into structured variables [27]. The corpus is designed to address challenges that arise in coding legal texts, including ambiguity in legal meaning and the way provisions interact within a document. DECODEM shares this concern with legally grounded text-to-variable translation, but differs in its focus on corporate organizational documents and in its evaluation of extraction pipelines for governance variables used in downstream empirical corporate-governance research. Other contract-focused benchmarks are also related insofar as they identify or classify provisions in contractual documents [22, 36].

Substantively, the closest existing benchmark is CHANCERY[23], which evaluates models on corporate governance reasoning tasks derived from real corporate charters. In CHANCERY, models are asked whether proposed executive, board, or shareholder actions are consistent with governance rules contained in the charter. The benchmark therefore shares this paper’s interest in corporate organizational documents and in the legal interpretation of governance provisions. The two projects nevertheless differ in important respects. CHANCERY is primarily a reasoning benchmark: it tests whether a model can apply governance rules in a charter to a proposed action. By contrast, the DECODEM benchmarks are designed to evaluate text-to-variable extraction. The task is not to reason about the permissibility of a hypothetical action, but to recover structured governance variables directly from primary documents. Relatedly, CHANCERY relies on charter-based scenarios constructed around governance principles, whereas the present paper focuses on document-level extraction tasks defined by hand-coded labels attached to randomly sampled real-world charters and bylaws.

More broadly, the paper relates to the literature on information extraction in natural language processing, which

studies how structured information can be recovered from unstructured text[31, 34]. Existing work in this area has developed a range of methods and benchmarks, often focusing on relatively well-defined extraction tasks applied to shorter or more standardized texts[10, 24, 37, 38]. By contrast, the extraction tasks considered here require document-level interpretation of lengthy and heterogeneous legal documents and the application of domain-specific concepts that are often not reducible to surface-level linguistic patterns[24].

Finally, the paper also relates to a rapidly growing literature examining the use of large language models for legal tasks. Recent work suggests that LLMs are capable of performing a wide range of such tasks, including legal reasoning and classification, question answering, and contract drafting (e.g., [6, 11, 28, 35]). A related strand of the literature highlights the ability of LLMs to extract structured information from legal sources and to generate high-quality summaries of judicial opinions and statutory materials[2, 7]. The present paper contributes to this literature by studying the automated recovery of structured variables from corporate governance documents in a setting designed to support large-scale empirical analysis.

Beyond the AI literature, the paper connects to the extensive law and finance literature studying the implications of corporate governance regimes (e.g., [4, 17]). In particular, it connects to recent work that has assembled large, manually collected datasets on corporate governance provisions in charters in order to examine whether central findings in empirical corporate governance are sensitive to improvements in data quality and measurement [14].

Even more broadly, the paper also relates to the broader literature on measurement in empirical social science. A central challenge in many empirical settings is the construction of reliable variables from complex source material, especially text[16, 18]. By enabling the automated recovery of legally meaningful variables from corporate charters and bylaws, the benchmarks and methods developed here have the potential to expand the set of governance features that can be studied at scale and reduce reliance on a limited set of legacy datasets.

5 The Benchmark Datasets

This paper introduces two novel benchmark datasets for evaluating the automated extraction of corporate governance information from the two central organizational documents of publicly traded corporations: charters and bylaws. Both DECODEM benchmarks share a common structure: each consists of a sample of documents paired with hand-coded labels capturing legally relevant governance arrangements, which define a set of document-level classification tasks used

for evaluation. The charters benchmark contains 300 charters, while the bylaws benchmark contains 150 bylaws.

Unlike many benchmark datasets in natural language processing, the tasks considered here are not just of interest as measures of model capability. Instead, they correspond directly to variables that have been used (or could plausibly be used) in empirical corporate governance research. As a result, improvements in extraction performance have immediate downstream consequences: they lower the cost of constructing high-quality datasets and make it possible to study governance features that have previously remained outside the reach of large-sample analysis.

The charters and bylaws were randomly selected from a comprehensive corpus containing the charters and bylaws of nearly all publicly traded corporations in the United States (see [12]). These documents were assembled through an automated pipeline using filings available on the SEC’s EDGAR database. The bylaws benchmark draws on documents filed between 1995 and 2024. For reasons related to the construction of the charters benchmark, the charters are drawn from the period between 1995 and 2019. Because both benchmarks are based on random samples from a corpus covering most charters and bylaws available on EDGAR, they support meaningful estimates of the out-of-sample accuracy of extraction pipelines designed to recover corporate governance provisions.

The bylaws benchmark covers 25 variables, the charters benchmark 26. Together, the benchmarks cover 31 distinct governance arrangements, implying that most variables feature in both datasets. This reflects the fact that U.S. corporate law often allows corporate planners to implement governance arrangements in either the charter or the bylaws. Table 1 summarizes all variables and reports their incidence in the bylaws and charters datasets. All variables take the form of binary indicators, which facilitates model evaluation. Some variables, for example WC-P/WC-A and SM-P/SM-A, address different aspects of the same governance topic. These variables are not complements, however, so including them as separate variables makes sense.

Together, the variables define a set of extraction tasks that vary in linguistic and interpretive complexity. Some variables (such as SB) primarily require the model to detect relatively standardized language, with little legally meaningful variation across documents. Others require more structurally complex interpretation of the document text. One example is ADJ, which requires determining whether the board has unilateral authority to adjourn a shareholder meeting and whether that authority is conditioned on the absence of a quorum. While the boundary between these categories is

Table 1: Variables Included in the Benchmark Datasets

Variable	Description	Count	
		BL	Ch
<i>Exculpation / Indemnification</i>			
IND-D	Indemnification of directors req.	108	81
IND-O	Indemnification of officers req.	107	77
ADV-D	Advancement req. for directors.	79	43
ADV-O	Advancement req. for officers.	79	41
INS-D	D&O insurance req. for directors.	6	NA
INS-O	D&O insurance req. for officers.	6	NA
EXC-D	Director exculpation.	3	153
EXC-O	Officer exculpation.	3	15
<i>Shareholder Meetings</i>			
SM-P	SHs cannot call special meetings.	75	52
SM-A	SHs can call special meetings.	50	22
SM-T*	Threshold support required.	49	17
WC-P	Written consent prohibited.	32	83
WC-A	Written consent allowed.	50	16
WC-S*	Supermajority required.	5	5
ADJ*	Unilateral adjournment.	49	NA
<i>Director Elections</i>			
REM*	Removal only for cause.	26	61
REM-S1*	Supermajority req. (for cause).	16	39
REM-S2*	Supermajority req. (w/o cause).	7	16
VAC-R*	SH vacancy authority (removal).	25	11
VAC-G*	SH vacancy authority (general).	11	5
MAJ	Majority voting.	35	11
PRA	Proxy access granted.	11	NA
CUMU	Cumulative voting.	NA	8
<i>Board Structure</i>			
BS-U*	Board authority to fix board size.	60	36
SB	Staggered board.	40	76
LID	Lead independent director.	8	NA
<i>Structural Features</i>			
BC	Blank check preferred.	NA	172
SMJ-C	Charter amendment smj. req.	NA	112
SMJ-M	Merger supermajority req.	NA	58
MC-U*	Multiclass shares w/ uneq. votes.	NA	27
TS-OO*	Anti-takeover statute opt out.	NA	19

Notes. Variables marked with an asterisk (*) require context-dependent or structurally complex interpretation, including the identification of conditional rules or numerical thresholds.

gradual rather than absolute, variables requiring more complex interpretation are marked with an asterisk in Table 1.

All variables were collected through a two-step process combining human coding with AI-assisted review. For the bylaws benchmark, most variables were coded first by a team of research assistants, all of them advanced law students at U.S.

law schools. Individual research assistants worked on only one variable, or a small set of closely related variables, at a time. For each variable, the same research assistant coded all bylaws in the benchmark dataset. Before coding began, research assistants received detailed instructions on how to interpret the variable they were assigned. Throughout the process, they were encouraged to flag ambiguities for review. A second set of variables, including most of the variables requiring advanced interpretation, was collected by the author.

The information collected at this stage was more detailed than the variables ultimately included in the benchmark dataset. In particular, for variables where the distinction was meaningful, coders recorded not only whether the bylaw contained the specific arrangement captured by the benchmark label, but also whether it addressed the broader governance topic at all. For example, one question asked coders to determine whether the bylaws contained any provision governing director removal. A second question then asked whether director removal was limited to removal for cause only. In addition, for all variables, coders were asked to quote the text on which they based their coding decision.

For the charters benchmark, most variables were obtained from a previous project in which we collected detailed information on the contents of charters with the help of a larger team of research assistants[14]. A small set of additional variables was collected by the author.

In a second step, the coding decisions and the previously available information were subjected to an AI-assisted audit scheme using earlier models than the one whose performance is evaluated later. The design of this audit differed depending on whether the available information indicated the presence or the absence of the relevant provision.

When the data indicated the presence of a relevant provision type, GPT-5.2 was prompted to review the corresponding coding decisions for individual variable categories—for example, the two questions related to director removal discussed above—against the text quoted by the human coder. Cases that GPT flagged as unsupported were reviewed either by the author or by a senior research assistant at Washington University in St. Louis, working under close supervision.

When the information collected in the first step indicated the absence of a relevant provision type, GPT-5-mini was prompted to review the full document text for paragraphs containing language related to the relevant governance topic. In instances where this model flagged potentially relevant language, GPT-5.2 was then prompted, in a second step, to assess whether the identified passages in fact contained the provision at issue. Cases for which the model answered the question in the affirmative were then reviewed by the author, who corrected the coding where necessary.

Two features of this process deserve emphasis because they ensure the suitability of the benchmarks for evaluating automated extraction methods. First, all final labeling decisions were made by humans. Automated methods were used only to flag coding decisions for review; in no case did an LLM produce the labels used to evaluate model performance. Second, the prompts used in the validation process were developed separately from the extraction routines evaluated in the following section. Both design choices enable a validation process that leverages automated methods while reducing the possibility of model-dependent biases.

Companion work extends this validation process one step further[13]. Every observation on which the initial human coding, the benchmark labels, and six of the extraction routines described below do not unanimously agree is recoded blindly, without knowledge of the existing labels. This adjudication upholds the benchmark labels in a large majority of reviewed cases, ultimately revising less than 1% of observations in the charters benchmark and around 1.5% of observations in the bylaws benchmark. This process produces an enhanced set of reference labels (the “Adjudicated Benchmarks”). Because the Adjudicated Benchmarks were constructed after observing disagreements between the DECODEM Benchmarks and the extraction routines evaluated here, this paper evaluates models and pipeline performance against the original DECODEM Benchmarks. This choice avoids using a reference set whose construction was partly triggered by the evaluated routines. Simultaneously, because some adjudicated revisions resolved model-benchmark disagreements in favor of the extraction routines, the performance estimates reported below are likely conservative.

The DECODEM benchmarks, the Adjudicated Benchmarks, the underlying charter and bylaw documents, the extraction prompts and pipelines, and the evaluation code will be made publicly available upon publication under a permissive license. Materials are available from the author on reasonable request prior to publication.

6 Model Performance—Bylaws

This section evaluates the performance of several automated extraction approaches on the bylaws benchmark. The goal is not only to identify which method performs best overall, but also to better understand how current extraction pipelines handle different types of governance provisions. Because the benchmark includes variables that vary in prevalence and linguistic complexity, the evaluation provides a first systematic picture of where automated extraction performs well and where substantial challenges remain. It also sheds

light on which architectural and design choices are associated with stronger extraction performance.

The extraction task consists of a series of document-level binary classification problems, one for each benchmark variable. Performance is evaluated by comparing model-generated labels to the benchmark labels described in the previous section.

I evaluate several extraction pipelines that vary along multiple dimensions, particularly in the amount of task-specific guidance they provide, the extent to which they operate on full documents rather than extracted passages, and the degree to which the task is decomposed into narrower, question-specific subtasks. By contrast, all pipelines instruct the model to return output in a fixed JSON schema that represents the benchmark variables in structured form and facilitates downstream processing.

The first set of approaches relies on simple prompts that ask the model whether a given governance arrangement is present in the text. These prompts do not provide detailed definitions of the relevant governance provisions or examples of how they may be expressed. Instead, they rely primarily on the model’s understanding of corporate governance terminology acquired during training. As a result, they require relatively little task-specific prompt engineering, making them a natural baseline for low-cost extraction. They also avoid the need to encode detailed domain knowledge directly in the prompt. For variable *SB*, the model is prompted as follows:

Detailed Questions:

– Staggered Board:

* `currently_staggered`: Does the bylaw stipulate that the corporation’s board is staggered/classified?

The second set of approaches relies on more detailed prompts that provide the model with background information and explicit rules for resolving harder cases. Unlike the simple prompts described above, these prompts do not rely primarily on the model’s latent understanding of corporate governance terminology. Instead, they supply task-specific guidance intended to sharpen distinctions between closely related provisions and to reduce errors in ambiguous cases. As a result, setting up these prompts requires substantially more domain-specific knowledge than in the case of simple prompts. For variable *SB*, the prompt includes the following questions and instructions:

Detailed Questions for JSON Fields:

– Contains Staggered Board Provision: Do the bylaws include any provision addressing staggered boards or the

division of directors into different classes? Answer "Y" or "N".

- Staggered Board:
 - * Staggered Board: Do the bylaws expressly stipulate that the board is staggered or that directors are divided into classes? "Y"/"N".

Scope/Interpretation:

- Contains Staggered Board Provision = "Y" if the excerpt discusses director classes in any sense (including share-class seat allocation), even if Staggered Board ultimately = "N" under the staggered-board test.
- Staggered/classified board test: Code Staggered Board="Y" only if the bylaws state or clearly imply that directors are divided into classes with multi-year terms and that only a subset of directors is elected at each annual meeting (e.g., "one class each year").
- Exclude share-class seat allocation: If "classes" language only allocates director election rights or board seats by classes/series of stock (e.g., "Class A elects X directors" or "Series A Preferred elects Y directors"), do not treat this as a staggered board.

In addition, the pipelines vary in whether the model is provided with the full bylaws or with selected excerpts. These excerpts are generated using a BERT- and LASSO-based pipeline that predicts, for each paragraph, whether it contains language relevant to a given governance provision (for details, see [12]). Because the downstream model can tolerate irrelevant passages, the excerpting step is designed to prioritize recall over precision. This approach can substantially reduce extraction costs by limiting the number of tokens submitted to the downstream classification model, while preserving the information needed to identify relevant provisions (see [26]). To the extent that model performance is sensitive to long context windows, this "cascading pipeline" may also improve performance.

Overall, I compare five extraction approaches that combine these design choices along three dimensions: the scope of extraction (global vs. topic-specific), the level of prompt

detail, and whether the model operates on full documents or selected excerpts.

Global: The full bylaw is provided to the model, which is asked—using the simple prompts—to extract all governance variables in a single pass.

Simple Full: The full bylaw is provided to the model, which is asked—using the simple prompts—to extract variables related to a specific governance topic.

Simple Cascade: Selected bylaw excerpts are provided to the model, which is asked—using the simple prompts—to extract variables related to a specific governance topic.

Detailed Full: The full bylaw is provided to the model, which is asked—using the detailed prompts—to extract variables related to a specific governance topic.

Detailed Cascade: Selected bylaw excerpts are provided to the model, which is asked—using the detailed prompts—to extract variables related to a specific governance topic.

Besides the extraction approach, I also vary the model used for the task, focusing on the following six current frontier and efficiency-oriented models from the three major AI labs: GPT-5.4, GPT-5.4-mini, Claude Opus 4.7, Claude Sonnet 4.6, Gemini Pro 2.5, and Gemini Flash 2.5. All models were accessed through their respective provider APIs between March 27, 2026 and June 19, 2026, using the following snapshots: gpt-5.4-2026-03-05, gpt-5.4-mini-2026-03-17, claude-opus-4-7, claude-sonnet-4-6, gemini-2.5-pro, and gemini-2.5-flash. For the models that allow for the setting of temperature, generation used temperature 0, with all other parameters left at provider defaults. To check for run-to-run variation, I reran the extraction up to three times for the principal models and approaches on both benchmarks. Variation was small enough that the following sections only report results from the first run.¹ Model outputs were constrained to the fixed JSON schema described above and parsed programmatically. Outputs that failed to parse or omitted a required field were re-queried under the same prompt and decoding settings until a valid JSON object was returned; this occurred only for formatting failures and not in response to the substantive content of a prediction. All prompts, extraction code, and evaluation scripts will be released with the benchmarks.

Model comparisons. Figure 1 compares performance across these six models for two extraction approaches: Global (left

¹For the frontier models, predictions were identical for roughly 99.5% of document-variable cases. Across all reruns, macro-averaged F1 (over variables with at least ten positive instances) changed by less than one point for every model except GPT-5.4-mini, which changed by about 1.4 points. Disagreements at the level of individual predictions were rare and concentrated in the interpretively complex variables that account for most extraction errors, and no substantive conclusion changes across runs.

column) and Detailed Cascade (right column). These approaches sit on two opposite sides of the spectrum of approaches included in this study. Global uses simple prompts and involves no preprocessing of the bylaws. Detailed Cascade uses the detailed prompts and relies on identifying relevant parts of a bylaw before the text is fed into the model. Within each approach, the darker box plots display variables requiring relatively little legal interpretation (the “Simple Variables”), while the lighter box plots represent variables requiring more context-dependent reasoning (the “Complex Variables”). To reduce distortions because of F1’s sensitivity to low-occurrence variables, Figure 1 (as well as Figures 4 and 5 below) includes only variables with Count ≥ 10 .

For the Simple Variables, performance is uniformly strong, with median F1 scores clustered close to one. This is true for both the Global and Detailed Cascade approaches. Complex Variables exhibit substantially greater variation. Within each model family, the larger models appear to outperform the more efficient ones. This is particularly true for the Global approach, where GPT-5.4-mini appears to perform noticeably worse than the frontier models. By contrast, there appears to be relatively little variation among the three frontier models under both approaches, with one small exception noted below.

To place these observations on firmer footing, I estimate paired bootstrap confidence intervals for differences in macro-averaged F1 between models.² Under the Global approach, the gap between frontier and efficiency-oriented models is clear: in each model family, the frontier model outperforms its smaller counterpart, with confidence intervals excluding zero. Under Detailed Cascade, this gap largely disappears. The GPT and Claude comparisons bracket zero, and for Gemini the ranking reverses, with Flash outperforming Pro by about one F1 point.

This pattern suggests an interaction between model capability and extraction architecture. For the frontier models, Global and Detailed Cascade yield broadly similar overall performance: the Global-to-Detailed-Cascade difference brackets zero for Opus 4.7 and Gemini Pro 2.5, GPT-5.4 improves marginally (0.04 to 3.2 F1 points). By contrast, Detailed Cascade substantially improves performance for the efficiency-oriented models, with gains distinguishable from zero in all three model families: roughly 7 to 12 F1 points for GPT-5.4-mini, 2 to 6 for Sonnet 4.6, and 1 to 4 for Gemini Flash. The corresponding difference-in-differences estimates, comparing

each efficiency-oriented model’s gain from Detailed Cascade to the gain for its frontier counterpart, also exclude zero in all three families. In short, the more elaborate pipeline mainly benefits the weaker models, lifting them toward frontier-level performance and, in the case of Gemini, slightly beyond it.

Among the three frontier models, pairwise differences are generally small. Under the Global approach, all three confidence intervals bracket zero and are no wider than roughly two F1 points. Under Detailed Cascade, GPT-5.4 outperforms Gemini Pro 2.5 by a small margin, with an interval excluding zero ([+0.005, +0.024]), while the other frontier-model comparisons continue to bracket zero. Because this is the only frontier comparison to exclude zero, its lower bound is close to zero, and the largest gap is only about two F1 points, I treat the three frontier models as effectively interchangeable for purposes of this task.

These results are notable. Even using simple prompts, current frontier models appear capable of extracting legally relevant information from complex legal documents such as corporate bylaws with a high degree of accuracy. At the same time, the figure shows that the lower tail in performance is driven almost entirely by variables requiring more complex interpretation. These results suggest that the main challenge is not average performance—which is near ceiling across all models—but variation across variables.

Figures 2 and 3 provide additional information about the comparative performance of the six models. Figure 2 visualizes how model errors are distributed across observations for each governance variable under the Global and Detailed Cascade approaches, here including the variables that were excluded before due to low N. Within each panel, observations are ordered from left to right by the number of models that misclassified them. The vertical axis indicates how many of the six models incorrectly classified a given observation. Flat lines near zero indicate variables for which all observations are classified correctly by all models, whereas increases toward the right side of a panel indicate a subset of observations on which one or multiple models fail.

Several patterns emerge. First, under both approaches, errors are highly concentrated among a relatively small subset of observations. For many variables, the curves remain close to zero for most of the distribution and rise sharply only near the right tail, indicating that most observations are classified correctly by nearly all models and that disagreements cluster around a comparatively small number of especially difficult cases.

Second, the shape of the curves differs systematically between the Global and Detailed Cascade approaches. Under the Detailed Cascade approach, the curves are often flatter for most of the distribution and then rise sharply at the ex-

²For each model pair, I draw 10,000 document-level bootstrap samples, recompute variable-level F1 scores for each model within each sample, and average the resulting differences across variables with at least ten positive instances. Intervals are computed separately for the Global and Detailed Cascade approaches using the 2.5th and 97.5th percentiles of the bootstrap distribution.

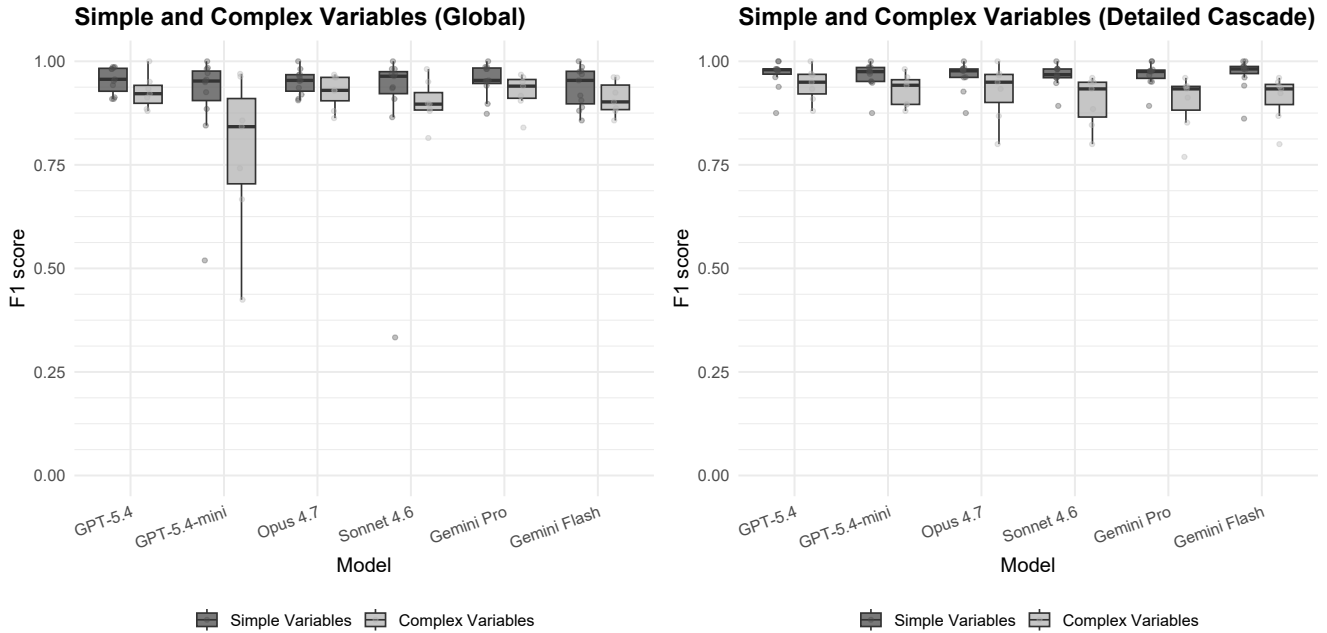


Figure 1: Variable-level F1 scores for extraction of bylaw features across models.

treme right tail. This suggests that models using Detailed Cascade tend to fail on a highly overlapping set of difficult observations. By contrast, the Global approach exhibits more gradual increases for a number of variables, including SM-P, WC-P, BS-U, and ADJ, indicating that model errors are more dispersed across observations and less tightly concentrated on a single shared set of difficult cases.

Figure 3 further explores this pattern by reporting pairwise Jaccard similarities between the error sets of different models under each approach. The Jaccard similarity between two models is defined as the proportion of shared errors among all observations misclassified by at least one of the two models. High values therefore indicate that models tend to fail on the same observations.

The results reveal a clear difference between the two extraction approaches. Under the Global approach, pairwise Jaccard similarities are substantially lower, generally ranging from approximately 0.25 to 0.55. This indicates that different models often make distinct mistakes when processing the full document context. Under Detailed Cascade, pairwise similarities range from approximately 0.36 to 0.71, and for every model pair, the cascade similarity exceeds its Global counterpart. This suggests that the cascade architecture induces more correlated failure patterns across models.

Taken together, these findings suggest that the Detailed Cascade approach induces substantially more correlated failure patterns across models than the Global approach. One possible interpretation is that the retrieval and decomposition stages of the cascade pipeline create shared informational bottlenecks: when the pipeline fails to surface or correctly contextualize relevant information, most frontier models tend to fail simultaneously. By contrast, the Global approach appears to leave more room for model-specific differences in long-context processing and legal interpretation, resulting in more heterogeneous error profiles across models.

Additional analyses comparing the error sets generated by the same model under the Global and Detailed Cascade approaches reveal only moderate overlap, with Jaccard similarities for all variables combined generally ranging from approximately 0.20 to 0.38. This suggests that the choice of extraction architecture affects not only overall performance levels but also which observations are misclassified.

Extraction approaches. Having compared performance across models, I now turn in more detail to variation across extraction architectures while holding the underlying model fixed. Figure 4 compares all five extraction architectures using GPT-5.4, one of the strongest performers for the task evaluated

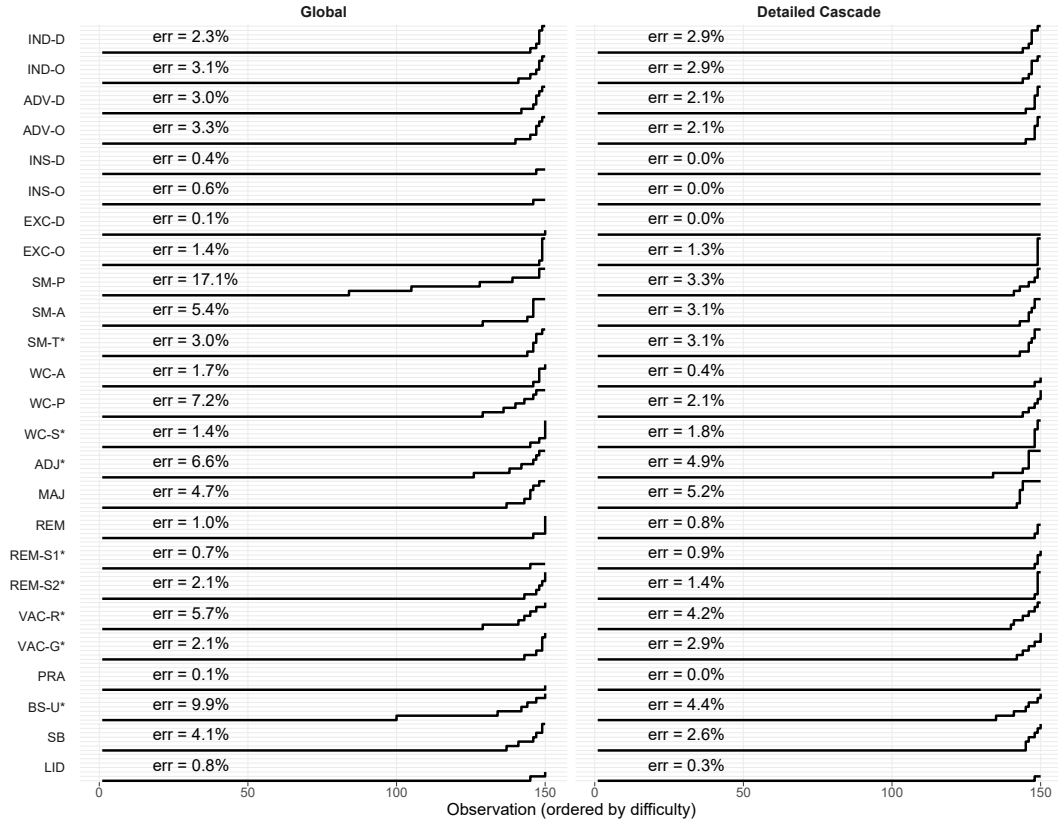


Figure 2: Lines represent the error concentration across models, with observations ordered by the number of models misclassifying each observation. *err* denotes the average error rate across models.

here. In the plot, the second row of boxplots displays the same distribution as the upper row but zooms in on the upper range.

Consistent with the results above, differences across approaches remain modest overall. Simple Full, Simple Cascade, and Detailed Full exhibit performance levels roughly in line with Global and Detailed Cascade. Comparing across approaches, neither cascading extraction nor more detailed prompting appears to consistently improve performance.

To support a cleaner comparison across approaches, I again estimate paired bootstrap confidence intervals, this time holding the model fixed and averaging F1 across variables with at least ten positive instances. For GPT-5.4, the five approaches fall within roughly two F1 points of one another. Decomposing the topic-specific approaches into prompt detail and document handling yields no contrast distinguishable from zero: the intervals for detailed prompting, excerpting, and their interaction all span zero. The same is true for the comparison between the Global approach and topic-specific full-document querying, which if anything slightly favors the

simpler single-pass approach. Repeating the same decomposition for Claude Opus 4.7 yields the same result.

Together, these results reinforce the conclusion that current frontier models recover these provisions about equally well across prompt designs, document-handling choices, and task decompositions. In particular, the fact that cascading approaches do not consistently outperform full-document approaches suggests that current frontier models can often utilize long legal contexts without extensive preprocessing.

Table 2 reports variable-level results for GPT-5.4. Several patterns emerge. First, a large set of variables—including indemnification and advancement provisions—can be extracted with very high accuracy across approaches. For these variables, F1 scores are uniformly close to one, and differences across approaches are negligible. A similar pattern holds for a number of less frequent variables, such as insurance, proxy access and lead independent director provisions, which also achieve near-perfect performance despite the small number of positive instances. What these variables share is that

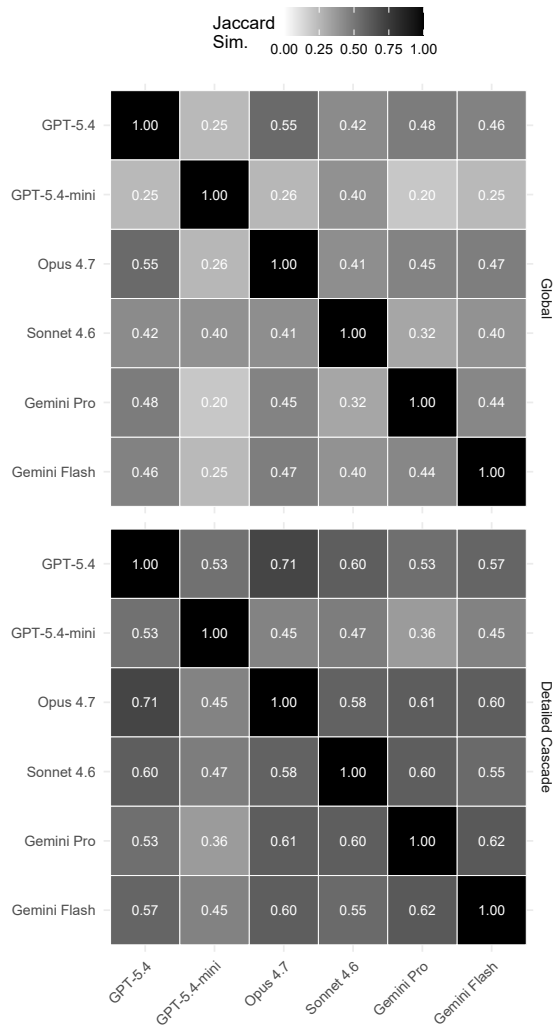


Figure 3: Pairwise Jaccard similarities between model error sets.

they tend to be expressed in relatively standardized language across bylaws, with little meaningful variation that would affect the ultimate coding decision. That said, indemnification, advancement, and insurance provisions vary in whether they require or merely allow the company to reimburse or provide insurance for the respective actor, and the models appear to correctly distinguish between both types of provisions.

Second, for a number of variables, performance remains high but exhibits more noticeable variation across approaches. This is particularly visible for the board structure provisions. In these cases, F1 scores remain close to one, but differences across approaches are more pronounced than for the near-

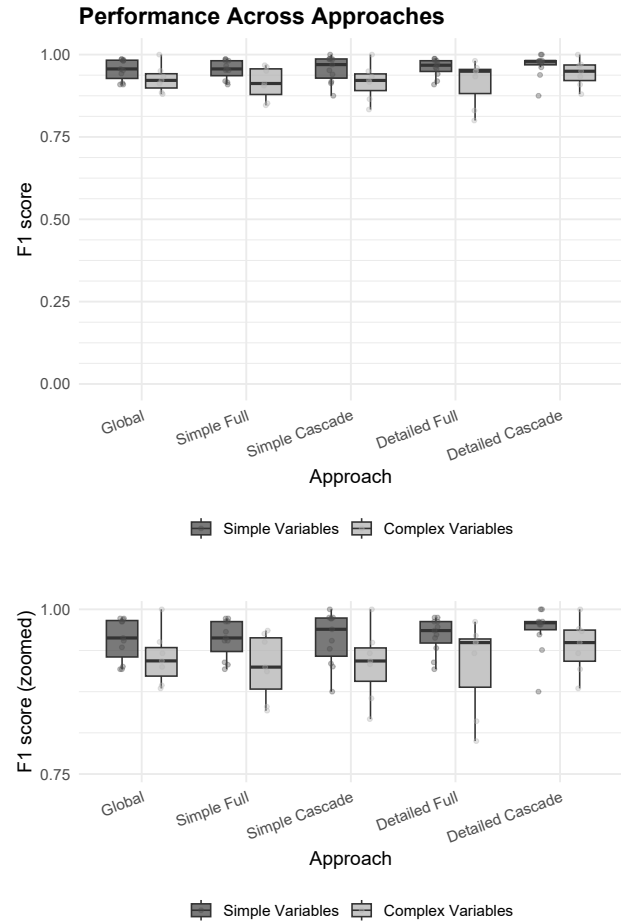


Figure 4: Variable-level F1 scores across extraction approaches. The top panel reports the full scale, while the bottom panel zooms in on the upper range.

uniform variables discussed above. Across these variables, however, no single approach consistently dominates: full-document approaches often perform well, but cascade-based approaches are best or tied for best for several provisions.

Third, the Complex Variables exhibit substantially greater variation across approaches and, on average, lower performance than the Simple Variables. Consider ADJ as an example. Here, performance differs markedly across approaches even when using the same simple prompting questions, while approaches based on more detailed prompting perform better. Conversely, for REM-S2, the simpler approaches outperform the detailed ones. These patterns illustrate that performance differences are not driven by a single superior approach, but

Table 2: Variable-Level Performance for GPT-5.4 on the Bylaws Benchmark

Variable	Count	Global	Simple Full	Simple Cascade	Detailed Full	Detailed Cascade
<i>Exculpation / Indemnification</i>						
IND-D	108	0.986	0.986	0.986	0.981	0.977
IND-O	107	0.986	0.986	0.986	0.986	0.977
ADV-D	79	0.981	0.981	0.988	0.988	0.981
ADV-O	79	0.981	0.981	0.988	0.988	0.981
INS-D	6	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0
INS-O	6	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0
EXC-D	3	3/0/0	3/1/0	3/3/0	3/0/0	3/0/0
EXC-O	3	1/0/2	1/1/2	1/3/2	1/0/2	1/0/2
<i>Shareholder Meetings</i>						
SM-P	75	0.913	0.966	0.913	0.973	0.980
SM-A	50	0.952	0.952	0.952	0.941	0.962
SM-T*	49	0.950	0.950	0.949	0.949	0.970
WC-P	32	0.985	0.952	0.970	0.968	0.980
WC-A	50	0.942	0.916	0.940	0.961	1.000
WC-S*	5	4/0/1	3/1/2	1/0/4	3/0/2	2/0/3
ADJ*	49	0.913	0.852	0.922	0.960	0.949
<i>Director Elections</i>						
REM*	26	1.000	0.963	1.000	0.981	1.000
REM-S1*	16	0.933	0.967	0.933	0.933	0.933
REM-S2*	7	7/0/0	7/0/0	7/0/0	5/0/2	5/0/2
VAC-R*	25	0.885	0.846	0.917	0.800	0.909
VAC-G*	11	0.880	0.906	0.833	0.830	0.880
MAJ	35	0.909	0.909	0.875	0.909	0.875
PRA	11	0.957	0.957	1.000	0.957	1.000
<i>Board Structure</i>						
BS-U*	60	0.922	0.912	0.865	0.950	0.967
SB	40	0.909	0.920	0.918	0.920	0.938
LID	8	8/1/0	8/1/0	8/0/0	8/0/0	8/0/0

Notes. Entries report F1 scores by variable and extraction approach for variables with Count ≥ 10 and TP/FP/FN (true positives / false positives / false negatives) for other variables. Count indicates the number of positive instances for the respective variable in the benchmark dataset. Variables marked with an asterisk (*) require context-dependent or structurally complex interpretation, including the identification of conditional rules or numerical thresholds. Performance for variables with very small counts should be interpreted with caution.

instead depend on the structure of the underlying provision and, potentially, on specific features of the prompts used.

Finally, performance is sometimes lower for relatively rare provisions, both in terms of their frequency in the dataset and their prevalence in practice. The comparison between EXC-D and EXC-O is instructive. These provisions often appear in the same bylaw documents, yet models are consistently better at detecting the former than the latter. One possible explanation is that officer-exculpating waivers were impermissible in most jurisdictions until very recently, while director-protecting waivers were standard in many charters[15]. As a result, models may internalize the expectation that exculpa-

tion provisions primarily protect directors, and fail to identify analogous provisions for officers even when they appear in bylaws.

Taken together, the results reinforce the high-level picture from Figure 4. Most governance provisions can be identified with high accuracy using relatively simple prompting strategies, while a limited set of variables (primarily those requiring more complex interpretation) accounts for most of the observed variation in performance, including the lower tail of the distribution. Differences across extraction approaches are generally modest, and no single method consistently dominates. At the same time, it is notable that approaches op-

erating on the full text often perform as reliably as cascade-based approaches. Despite the substantial length of bylaw documents, which often exceed 10,000 words and sometimes surpass 20,000, providing the model with the full text does not appear to impair performance. If anything, in some cases, it appears to facilitate interpretation by allowing the model to situate individual provisions within the broader structure of the document.

Additional analyses further suggest that raw document length does not impair performance. Although longer bylaws initially appear associated with somewhat higher error rates, this relationship disappears once the analysis controls for whether a filing is a full restatement rather than a partial amendment that contains only selected bylaw provisions. Partial amendments exhibit substantially lower error rates than full restatements. Yet they are not only shorter than full restatements, they also typically address only a small subset of governance provisions and are therefore structurally less complex. By contrast, among full restatements, there appears to be no meaningful relationship between document length and extraction errors despite substantial variation in document length, further strengthening the conclusion that context window size does not appear to pose a significant obstacle for current frontier models in this type of task.

7 Model Performance—Charters

This section extends the evaluation to the charters benchmark, testing whether the high-accuracy findings reported for bylaws generalize to a second class of organizational document. Charters differ from bylaws in several ways. They are often shorter, which might facilitate extraction; at the same time, their structure can vary substantially, in particular because charters may contain detailed provisions governing preferred stock. Throughout, I rely on the same extraction pipelines, models, and document-level binary-classification framing introduced in the previous section, so that results are directly comparable across the two document types.

Figure 5 replicates Figure 1, reporting variable-level F1 scores across the six models for the Global and Detailed Cascade approaches and again excluding variables with a count of < 10 . The overall picture broadly mirrors the bylaws results. For the Simple Variables, performance is uniformly strong, with median F1 scores clustered near one under both approaches. The Complex Variables exhibit substantially greater dispersion, and, as in the bylaws analysis, the lower tail of the distribution is driven almost entirely by these variables. Within each model family, the larger models tend to outperform their efficiency-oriented counterparts on the Com-

plex Variables, a gap most visible for GPT-5.4-mini under the Global approach.

Bootstrap comparisons, however, reveal two departures from the bylaws benchmark. First, the frontier models are less interchangeable here: under the Global approach Opus 4.7 outperforms the other two by intervals excluding zero, though the three converge under Detailed Cascade. Second, the model-size $\times$ approach interaction observed for the bylaws benchmark only replicates for the GPT model family. Only GPT-5.4-mini gains systematically under Detailed Cascade, and some models, notably Opus 4.7, perform worse under Detailed Cascade than under Global.

Furthermore, comparing Figures 5 and 1 suggests that average performance on the charters benchmark trails that on the bylaws benchmark.

Figure 6 sheds further light on this gap by comparing the variables that appear in both benchmarks. The figure plots average F1 for the three frontier models on the charters benchmark (y-axis) against average F1 on the bylaws benchmark (x-axis) under the Global and Detailed Cascade approaches; the dashed 45-degree line marks equal performance. Points below the line correspond to variables for which extraction from charters trails extraction from bylaws, while points above the line mark variables recovered more accurately from charters. Colors distinguish variables with ≥ 10 observations in both benchmarks (those variables included in Figures 1 and 5) from variables for which this condition is not met.

Because these variables are extracted using identical prompts, their placement helps distinguish document-type effects from differences in variable selection and prompt design. Systematic underperformance on charters would suggest that some structural feature of charter documents makes legally relevant information harder to recover, whereas placement close to the line would indicate that the strong bylaws results are not related to the specific structure of bylaws.

As the figure shows, most shared variables cluster in the upper-right corner, close to the diagonal, indicating that extraction performance is broadly stable across document types: provisions that are easy to recover from bylaws are, for the most part, equally easy to recover from charters. A few variables depart from this pattern.

The first group of variables with a different pattern belongs to the variables with low N in at least one of the benchmarks. EXC-O, for example, lies well above the line, reflecting markedly better performance on charters than on bylaws. The most likely explanation is statistical: the bylaws benchmark contains only three positive instances, so its F1 score is highly sensitive to individual errors, whereas the larger number of positive instances in the charters benchmark yields a more stable estimate. A secondary possibility is that models more

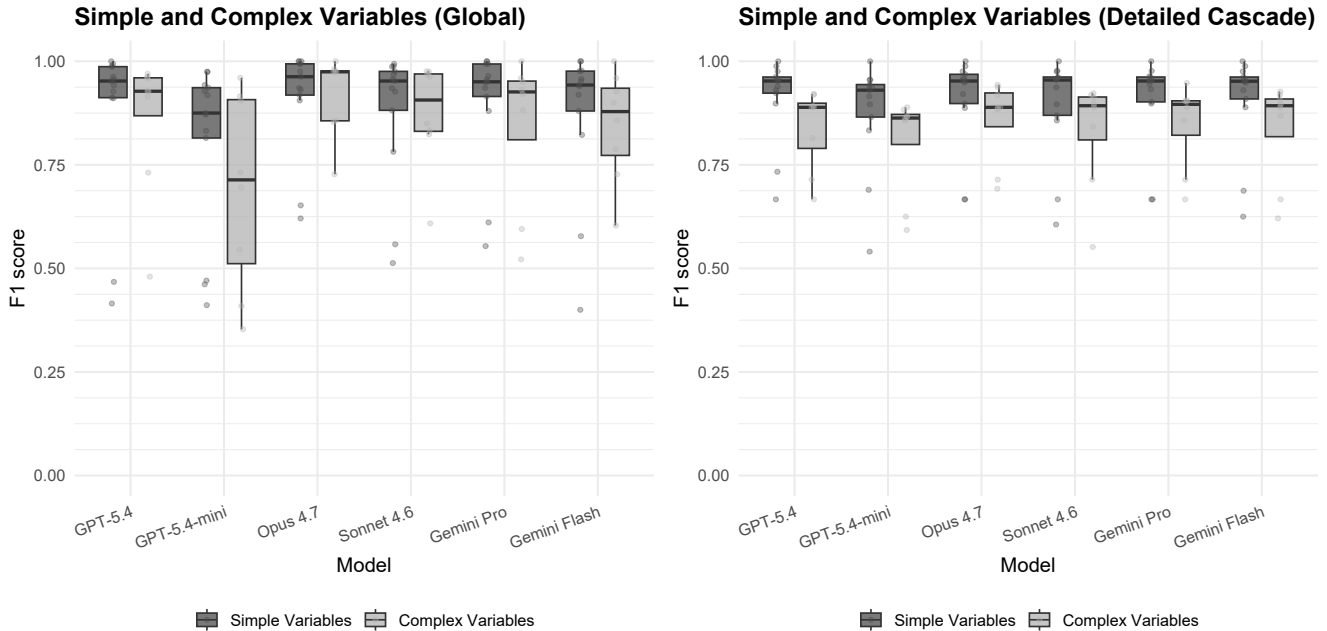


Figure 5: Variable-level F1 scores for extraction of charter features across models.

readily “expect” to encounter officer-exculpation provisions in charters than in bylaws, but this possibility is impossible to verify given the low count of true “Ys” for bylaws.

Another group of variables consists of variables with sufficient support in both benchmarks. These variables include VAC-R and various variables from the WC and SM categories. All these variables exhibit stronger performance in the bylaws benchmark. This difference in performance is likely due to the distorting effect of provisions governing preferred stock, which sometimes contain provisions applicable exclusively to the holders of specific series of preferred stock. Those provisions were excluded from the definitions used to assemble the benchmarks but likely confuse the extraction algorithms in some instances.

To corroborate the document-type gap visible in Figure 6, I compute bootstrap confidence intervals for the difference in macro-averaged F1 between bylaws and charters on variables with at least ten positive instances in both benchmarks, matching on variables while resampling the two benchmarks independently. Across the Global and Detailed Cascade approaches and the three frontier models, bylaws outperform charters by roughly 5 to 11 F1 points, with all confidence intervals excluding zero. Because the comparison holds the variable set and prompts fixed, the gap is not an artifact of variable

selection but reflects differences associated with extracting the same governance variables from the two document types.

In sum, the charters results support the central conclusions of the bylaws analysis. Automated extraction is feasible at a high level of accuracy for the large majority of governance provisions, including several structural and anti-takeover features that do not appear in the bylaws benchmark. Performance again varies systematically across variables, with interpretively complex provisions accounting for most of the remaining errors, and more detailed prompting and cascading pipelines again do not deliver consistent gains for frontier models. The close correspondence between the two benchmarks shown in Figure 6 provides additional evidence that these findings are broadly consistent across document types, with a few genuine exceptions concentrated in shareholder-meeting, written-consent, and vacancy provisions, plausibly from preferred-stock interference in charters.

8 Limitations

Several limitations apply. Most importantly, the benchmark labels likely contain some noise, and they were not produced through a process that would permit the computation of inter-coder reliability. This noise places a ceiling on measured accuracy: an apparent model error may in some

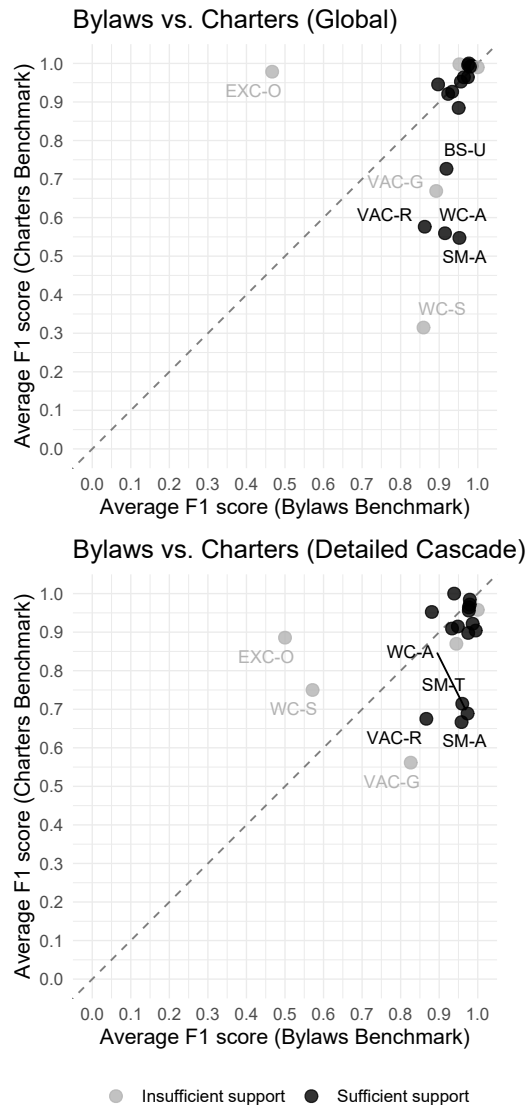


Figure 6: Comparison of performance for variables featured in both benchmarks. Average F1 scores are computed across GPT-5.4, Claude Opus 4.7, and Gemini Pro 2.5.

instances reflect a borderline coding decision or coder mistake rather than a genuine extraction failure. The AI-assisted audit described in Section 5 was designed to surface and correct such errors, but it does not substitute for independent double-coding, and the labels should be understood as high-quality but not error-free. The blind re-coding exercise in Frankenreiter [13] offers a partial measure of the remaining noise: it revises less than 1% of charter observations and around 1.5% of

bylaw observations, suggesting that residual error, while real, is unlikely to drive the aggregate results. Besides, robustness checks suggest that the results reported here do not materially change if the Adjudicated Benchmarks are used in place of the DECODEM Benchmarks. Still, because some adjudicated revisions resolved model-benchmark disagreements in favor of the extraction routines, the accuracy estimates reported here are likely conservative.

Second, the benchmarks embed interpretive choices about what each variable captures, and those choices affect measured difficulty. Most notably, provisions applicable exclusively to holders of specific series of preferred stock were excluded from the variable definitions; as discussed in Section 7, this exclusion likely contributes to the weaker performance on several shareholder-meeting and written-consent variables in the charters benchmark, where these provisions appear. More generally, the measured difficulty of a variable reflects in part where its boundary was drawn, and a different but equally defensible coding rule could shift the results.

The benchmarks are also bounded in coverage. The documents are drawn from a corpus of SEC EDGAR filings for U.S. publicly traded corporations (see [12]); the bylaws span filings between 1995 and 2024, while the charters cover 1995 through 2019. The estimates therefore speak to that population and those periods, and may not extend to private firms, non-U.S. organizational documents, or charters filed after 2019, particularly given that drafting conventions evolve over time.

Relatedly, because hosted frontier models change over time, the reported performance comparisons should be understood as a time-stamped evaluation of the model snapshots listed in Section 6.

Finally, two narrower points deserve mention. The first is data contamination. The charters and bylaws used here are public filings that have been available on EDGAR for years, so some may have appeared in the models' pretraining data. This is difficult to rule out and could mean that a portion of the measured accuracy reflects familiarity with specific documents rather than a general capacity to extract governance provisions. That said, the task requires applying a specific coding rule rather than reproducing a document, and, with the exception of some variables in the charters benchmark, the specific coding scheme used here does not correspond to any published dataset the models could have seen during training. Even so, the estimates are best read as accuracy on documents of this kind and vintage, and performance on genuinely novel filings could differ. Second, label construction relied on an AI-assisted audit using earlier OpenAI models (GPT-5.2 and GPT-5-mini), while the evaluation includes a more recent model from the same family among others. Although all final

labels were set by human coders and the validation prompts were developed separately from the extraction routines, the possibility of a mild same-family advantage cannot be entirely excluded.

9 Conclusion

This paper introduces the DECODEM benchmarks and uses them to evaluate the performance of large language models in extracting corporate governance variables from charters and bylaws. The results show that automated extraction is feasible at a high level of accuracy for many commonly studied provisions. At the same time, performance varies systematically across variables. Remaining errors are concentrated in tasks such as shareholder meeting rights, vacancy-filling authority, and rare officer-exculpation provisions, rather than in standardized indemnification or insurance provisions. Importantly, more elaborate prompting strategies and cascading extraction pipelines do not consistently improve performance for frontier models, but they substantially narrow the gap between frontier and efficiency-oriented models in some settings. This suggests that pipeline design can partly substitute for model capability, while the main obstacle to further gains for the strongest models lies less in extraction architecture than in the interpretive difficulty of particular provisions.

These findings have two main implications. First, they demonstrate that large language models can substantially reduce the cost of constructing governance datasets, potentially enabling empirical research at a scale that was previously infeasible. Second, they highlight the importance of careful task design and validation: even when average performance is high, certain provisions remain difficult to extract and require particular attention in downstream applications.

More broadly, the paper contributes to a growing literature on the use of artificial intelligence in legal research by shifting the focus from legal reasoning tasks to the measurement of legal institutions. Many questions in empirical legal studies ultimately depend on the availability of reliable variables derived from complex source material. By showing that such variables can be recovered with a high degree of accuracy from corporate governance documents, the paper points to a path toward expanding the empirical study of law beyond the relatively small set of features available in existing datasets and without the need for resource-intensive hand coding.

References

- [1] Mohamed Afane, Emaan Hariri, Derek Ouyang, and Daniel E. Ho. *Benchmarking Legal RAG: The Promise and Limits of AI Statutory Surveys*. 2026. arXiv: 2603.03300.
- [2] S Ajay Mukund and K. S. Easwarakumar. “Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation”. In: *Symmetry* 17.5 (2025).
- [3] Farid Ariai, Joel Mackenzie, and Gianluca Demartini. “Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges”. In: *ACM Computing Surveys* 58.6 (2025), pp. 1–37.
- [4] Lucian Bebchuk, Alma Cohen, and Allen Ferrell. “What Matters in Corporate Governance?” In: *The Review of Financial Studies* 22.2 (2009), pp. 783–827.
- [5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL]. URL: <https://arxiv.org/abs/2004.05150>.
- [6] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. “Can GPT-3 Perform Statutory Reasoning?” In: *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL ’23)*. ACM, 2023, pp. 22–31. DOI: 10.1145/3594536.3595163.
- [7] J. Breton, M. M. Billami, M. Chevalier, et al. “Leveraging LLMs for Legal Terms Extraction with Limited Annotated Data”. In: *Artificial Intelligence and Law* (2025).
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.
- [9] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. “LexGLUE: A Benchmark Dataset for Legal Language Understanding in English”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 4310–4330.

- [10] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. “The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation”. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*. Lisbon, Portugal: European Language Resources Association (ELRA), 2004.
- [11] Christoph Engel and Richard H. McAdams. “Asking GPT for the Ordinary Meaning of Statutory Terms”. In: *University of Illinois Journal of Law, Technology and Policy* (2024), pp. 235–296.
- [12] Jens Frankenreiter. *The Other Delaware Effect*. Washington University in St. Louis Legal Studies Research Paper No. 25-03-11. 2025. URL: <https://ssrn.com/abstract=5115285>.
- [13] Jens Frankenreiter. “Measuring Corporate Governance with Large Language Models”. 2026. URL: <https://ssrn.com/abstract=7132098>.
- [14] Jens Frankenreiter, Cathy Hwang, Yaron Nili, and Eric Talley. “Cleaning Corporate Governance”. In: *University of Pennsylvania Law Review* 170 (2021), pp. 1–70.
- [15] Jens Frankenreiter and Eric Talley. “Sticky Charters? The Surprisingly Tepid Embrace of Officer-Protecting Waivers in Delaware”. In: *Harvard Business Law Review* (2026). Forthcoming.
- [16] Matthew Gentzkow, Bryan Kelly, and Matt Taddy. “Text as Data”. In: *Journal of Economic Literature* 57.3 (2019), pp. 535–574. DOI: 10.1257/jel.20181020.
- [17] Paul Gompers, Joy Ishii, and Andrew Metrick. “Corporate Governance and Equity Prices”. In: *The Quarterly Journal of Economics* 118.1 (2003), pp. 107–156.
- [18] Justin Grimmer and Brandon M. Stewart. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. In: *Political Analysis* 21.3 (2013), pp. 267–297. DOI: 10.1093/pan/mps028.
- [19] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chonlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, et al. “LEGALBENCH: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models”. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. 2023, pp. 44123–44279.
- [20] Mark A. Hall and Ronald F. Wright. “Systematic Content Analysis of Judicial Opinions”. In: *California Law Review* 96.1 (2008), pp. 63–122.
- [21] Emaan Hariri and Daniel E Ho. “AI for Statutory Simplification: A Comprehensive State Legal Corpus and Labor Benchmark”. In: *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*. ICAIL ’25. New York, NY, USA: Association for Computing Machinery, 2026, pp. 177–187.
- [22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. “CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review”. In: *NeurIPS Datasets and Benchmarks* (2021).
- [23] Lucas Irwin, Arda Kaz, Peiyao Sheng, Sewoong Oh, and Pramod Viswanath. *CHANCERY: Evaluating Corporate Governance Reasoning Capabilities in Language Models*. 2025. arXiv: 2506.04636 [cs.AI]. URL: <https://arxiv.org/abs/2506.04636>.
- [24] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 6, 2026. 2026. URL: <https://web.stanford.edu/~jurafsky/slp3/>.
- [25] Kobi Kastiel and Yaron Nili. “The Corporate Governance Gap”. In: *Yale Law Journal* 131 (2022), pp. 782–857.
- [26] Zehua Li, Neel Guha, and Julian Nyarko. “Don’t Use a Cannon to Kill a Fly: An Efficient Cascading Pipeline for Long Documents”. In: *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*. 2023, pp. 141–147.
- [27] Florencia Marotta-Wurgler and David Stein. “Building a Long Text Privacy Policy Corpus with Multi-Class Labels”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. July 27–August 1, 2025. Association for Computational Linguistics, 2025, pp. 8156–8219.
- [28] Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. *Better Call GPT, Comparing Large Language Models Against Lawyers*. 2024. eprint: 2401.16212. URL: <https://arxiv.org/abs/2401.16212>.
- [29] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. “Distant supervision for relation extraction without labeled data”. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 1003–1011.

- [30] Julian Nyarko. “Stickiness and Incomplete Contracts”. In: *University of Chicago Law Review* 88.1 (2021), pp. 1–79.
- [31] D. Premasiri, T. Ranasinghe, R. Mitkov, et al. “Survey on Legal Information Extraction: Current Status and Open Challenges”. In: *Knowledge and Information Systems* 67 (2025), pp. 11287–11358.
- [32] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. “Snorkel: Rapid Training Data Creation with Weak Supervision”. In: vol. 11. 3. 2017, pp. 269–282.
- [33] Gabriel Rauterberg and Eric Talley. “Contracting Innovation”. In: *Columbia Law Review* 117.8 (2017), pp. 1701–1764.
- [34] Sunita Sarawagi. “Information Extraction”. In: *Foundations and Trends in Databases* 1.3 (2008), pp. 261–377.
- [35] Hannah Sargeant, A. Izzidien, and Felix Steffek. “Topic Classification of Case Law Using a Large Language Model and a New Taxonomy for UK Law: AI Insights into Summary Judgment”. In: *Artificial Intelligence and Law* (2025).
- [36] Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. “LEDGAR: A Large-Scale Multilabel Corpus for Text Classification of Legal Provisions in Contracts”. In: *Proceedings of the 12th Conference on Language Resources and Evaluation*. Marseille: European Language Resources Association, 2020, pp. 1235–1241.
- [37] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*. 2019, pp. 3266–3280.
- [38] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 353–355.
- [39] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. “Big Bird: Transformers for Longer Sequences”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 17283–17297.