JGovCCC-PDL: Japanese Government Contract Clause Corpus based on Public Data License

Nobushige Doi
The University of Tokyo
Tokyo, Japan
n-doi@g.ecc.u-tokyo.ac.jp

Takehisa Yairi
The University of Tokyo
Tokyo, Japan
yairi@g.ecc.u-tokyo.ac.jp

Abstract

We present JGovCCC-PDL (Japanese Government Contract Clause Corpus based on Public Data License), a Japanese corpus annotated for clause-type classification in public procurement contracts. We collected 212 publicly available contract documents (e.g., standard contracts and terms and conditions) from 32 public institutions, including the Japanese national government and local municipalities. All documents were released under the Japanese Public Data License (Version 1.0) or comparable terms that permit reproduction, public transmission, and adaptation with appropriate attribution. We segmented the documents into 9,281 clauses and annotated each clause with two label levels: a fine-grained clause-type label (50 classes) and a coarse-grained category label (9 classes). We further define a leakage-resistant evaluation protocol tailored to duplicate-heavy procurement contracts. Because clause-type identification is a prerequisite for downstream contract review, clause retrieval, issue extraction, and later contract reasoning, we position JGovCCC-PDL as a foundational resource for Japanese contract-oriented Legal NLP.

CCS Concepts

• **Applied computing** → **Law**; • **Computing methodologies** → **Language resources**; *Supervised learning by classification*; *Natural language processing*.

Keywords

Japanese contract, Legal provision annotation, Corpus construction

1 Introduction

Legal natural language processing (Legal NLP) applies natural language processing (NLP) to legal texts such as contracts, statutes, and case law, and supports tasks such as information extraction, classification, retrieval, and summarization [1, 3, 7, 11, 14]. Despite this progress, the limited availability of publicly reusable datasets continues to hinder reproducible evaluation and cross-study comparison, especially for contracts [6, 11, 13, 17]. In contract practice, clauses are a primary unit for drafting and review; accordingly, clause-level quantitative analysis and automation are important for both practitioners and researchers [5, 17, 19].

Public procurement contracts are often published for administrative transparency, and standard templates and terms accumulate across procurement types (e.g., construction, goods, and services). However, wording and structure vary across institutions and domains, so clause-level comparison, search, issue organization, and risk identification still rely heavily on manual work. In Japanese Legal NLP, redistributable clause-level corpora remain scarce, limiting fundamental resources for clause-type classification.

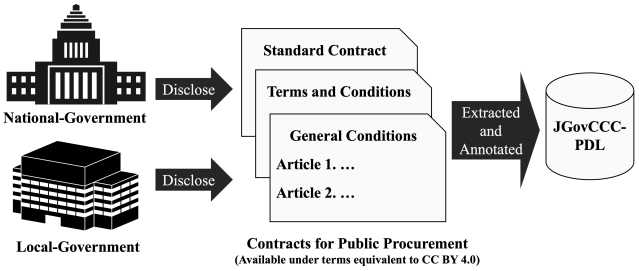

Figure 1: Overview of JGovCCC-PDL.

Clause-type classification is not an end in itself. In practical contract-analysis pipelines, AI systems must first identify what kind of clause they are processing before they can retrieve analogous provisions, compare clauses across templates, organize issues, flag risk-related provisions, or apply clause-specific reasoning. We therefore position clause classification as a foundational task for contract-oriented AI and law rather than as an isolated benchmark.

To address this gap, we present JGovCCC-PDL (Japanese Government Contract Clause Corpus based on Public Data License)[1], a Japanese corpus of clauses from public procurement contract documents. We collected 212 documents (e.g., standard contracts, terms and conditions, general conditions, and regulations) published by 32 public institutions, including the Japanese national government and local municipalities, and segmented them into 9,281 clauses. Each clause is annotated with a two-level label hierarchy: a fine-grained clause-type label (50 classes) and a corresponding coarse-grained category label (9 classes). To support redistribution and downstream research, we restricted sources to documents released under licenses that permit reproduction, public transmission, and adaptation with attribution, such as the Japanese Public Data License (Version 1.0)[2] [4] and terms equivalent to the Creative Commons Attribution 4.0 International (CC BY 4.0.) license. Figure 1 shows an overview of JGovCCC-PDL.

We also provide scripts and a standardized experimental protocol, including: (i) duplicate and label-conflict checks based on exact matching of clause text, (ii) subcorpus design using frequency thresholds and related criteria, (iii) data splits that control leakage, and (iv) bidirectional transfer evaluation between national-government and local-government subsets.

Our contributions are:

- We release JGovCCC-PDL, a redistributable Japanese clause corpus for public procurement contracts (212 documents;

[1] https://github.com/n-doi/JGovCCC-PDL
[2] https://www.digital.go.jp/en/resources/open_data/public_data_license_v1.0

9,281 clauses) annotated with 50 fine-grained clause types grouped into 9 coarse categories.
- We define a leakage-resistant evaluation protocol for this domain, combining duplicate-aware evaluation with `national` ↔ `local` transfer settings motivated by extensive template reuse.
- We provide reproducible baseline results using standard classical and Transformer-based models [18]. The contribution is the resource and benchmark rather than a new model architecture; the baselines are intentionally conventional so that future work can compare against a transparent starting point.

2 Related Work

2.1 Corpus of English Contracts

Public English contract corpora have enabled more standardized, reproducible evaluation in Legal NLP, including benchmark-style comparisons [3, 11]. For clause/provision-type classification, the LEDGAR corpus is a widely used large-scale dataset built from SEC filings [17]. For contract review, CUAD provides expert span annotations over a fixed set of issue types [6], while MAUD targets M&A agreements with annotations aligned to deal-point questions [19]. Beyond extraction and classification, ContractNLI frames contract understanding as document-level NLI with supporting evidence [8], and CLAUDETTE focuses on identifying potentially unfair clauses in online terms of service [9]. While these resources cover diverse sources and annotation targets, comparable reusable resources remain limited for Japanese contract clauses.

2.2 Research in the Japanese Legal Field

Japanese Legal NLP has primarily emphasized statutes and case law. COLIEE provides recurring shared tasks for legal retrieval and entailment/QA, supporting community-wide benchmarking [7, 12]. For judicial documents, corpora have been developed for structural and argument-oriented analysis [21] and for rationale-supported legal judgment prediction in tort cases [22]. Studies on Japanese legal pre-trained language models also highlight the practical limitations imposed by the scarcity of large, publicly available legal corpora [10]. Contract-focused Japanese resources are comparatively fewer; Funaki et al. introduce a bilingual contract corpus with annotations related to deontic and agent information [5]. JBE-QA expands Japanese evaluation resources via bar-exam QA, but it is oriented toward statutory knowledge rather than contract drafting and review [2]. These gaps motivate the creation of a redistributable Japanese clause corpus for public procurement contracts.

3 Corpus Construction

This section describes the construction of JGovCCC-PDL (Japanese Government Contract Clause Corpus based on Public Data License): (i) data sources and licensing criteria, (ii) clause segmentation and preprocessing, (iii) label design and annotation, and (iv) duplicate/conflict checks and corpus statistics.

Table 1: Size of JGovCCC-PDL by domain.

Domain	#Clauses	#Docs	#Orgs	#Labels
`national`	6,577	156	22	50
`local`	2,704	56	10	47
All	9,281	212	32	50

3.1 Target Documents and Collection Policy

JGovCCC-PDL targets standardized public procurement contract documents (e.g., standard contract templates and standard terms and conditions) published by the Japanese national government and local municipalities. To reduce the risk of including sensitive or personally identifiable information, we focused on template-style documents rather than project-specific contracts.

To ensure redistributability, we restricted sources to documents that explicitly permit reproduction, public transmission, and adaptation with appropriate attribution (e.g., source and license statements). Specifically, we included documents released under the Japanese Public Data License [4] or comparable terms that allow copying, modification, and redistribution with attribution.

3.2 Collection Scope and Metadata Design

We collected 212 documents from 32 public institutions (national: 22; local: 10). For analysis and evaluation, we partition the corpus into `national` and `local`. After segmentation, the corpus contains 9,281 clauses in total. Table 1 summarizes corpus size by domain.

The collected documents are not executed one-off contracts but reusable standard forms and conditions. They cover several procurement families, including public works and construction, architectural and civil-engineering design or supervision, goods purchase and manufacturing, service outsourcing and maintenance, system development and operation, forestry works and timber sales, leasing and rental, labor dispatch, industrial-waste collection or disposal, and research/content contracts containing Bayh-Dole-style intellectual-property clauses. Thus, although the corpus is limited to public procurement, it includes clauses on performance management, inspection and delivery, payment, contract change, termination, damages, confidentiality, personal information and system access, intellectual property, anti-social-forces exclusion, and environmental measures.

Each clause record stores the clause text together with document- and organization-level identifiers (including a domain flag) and structural fields such as article number and heading, enabling splits by document, institution, and domain.

3.3 Clause Segmentation and Preprocessing

We segmented each document into clauses based on its structure (article numbers, headings, and paragraph boundaries), with manual corrections when necessary. Because headings often vary in surface form (e.g., parenthesized expressions, symbols, whitespace), we store both the raw heading and a normalized version to support matching and analysis.

Text normalization is intentionally minimal to preserve legal meaning and support reproducibility. We apply simple formatting

Table 2: Clause counts and number of fine-grained labels by coarse category.

Coarse category (lv1)	#Clauses	No. of lv2 labels
Performance and Administration	2,392	11
Warranties, Liability, and Sanctions	1,433	4
Contract Termination and Disputes	1,404	7
Change and Risk	1,172	7
Money and Accounting	1,018	6
Rights and Information	930	4
General Terms and Communication	642	7
Compliance	193	3
Other	97	1

operations such as converting full-width spaces to half-width, collapsing consecutive whitespace, unifying newline codes, and compressing excessive consecutive newlines. With heading normalization, the number of unique headings decreased from 1,236 (raw) to 1,153 (normalized).

3.4 Label Scheme and Annotation

JGovCCC-PDL uses a two-level label hierarchy. Each clause is assigned a single fine-grained clause-type label (`lv2`; 50 classes), and the corresponding coarse-grained category label (`lv1`; 9 classes) is determined deterministically via a mapping table from `lv2` to `lv1`. The label inventory was designed through discussion among annotators, each with at least three years of practical experience in public works or contract administration. Annotation proceeded in two stages. First, the annotators jointly reviewed the full corpus and discussed recurring clause functions to define the fine-grained label inventory and its mapping to `lv1`. Second, they annotated the clauses using this scheme. When ambiguous cases or conflicts arose during annotation, they were resolved through discussion until a single final label was agreed upon for the released corpus. Because the public release stores only adjudicated final labels, we do not report a separate inter-annotator agreement coefficient in this paper. Table 2 shows the distribution by `lv1`.

The label distribution is skewed but not extreme: the most frequent label is *Damages and Liability* (675 clauses), and the least frequent is *Definitions* (12 clauses). All 50 labels appear in at least five documents; therefore, frequency thresholding mainly affects long-tail clause types.

Table 3 provides representative examples from the released corpus. The examples show that the labels do not merely separate broad contract domains; they capture clause functions that recur across different document families. For instance, an inspection clause may appear in a maintenance-service form, an access-control clause in a system-operation form, and an intellectual-property-use clause in a research or content contract.

3.5 Quality Control

Because boilerplate clauses are widely reused across public institutions, we detect duplicates and potential label conflicts based on exact matches of minimally normalized clause body text (whitespace/newline normalization). We found 3,662 duplicate clause bodies (counting each repeated occurrence beyond the first), meaning that 39.5% of clauses are reoccurrences of identical text (5,619 unique clause bodies). Duplicate bodies formed 1,216 groups, and the largest group size was 50. Most duplicate groups span multiple documents. We did observe a small number of within-document exact duplicates: 12 duplicate pairs (24 clauses in total), all consisting only of the placeholder text *deleted* and labeled as *Other*. Excluding these placeholder cases, duplicate reuse is overwhelmingly cross-document.

We observed zero cases of label-conflict duplicates (the same clause body assigned multiple labels), suggesting consistent labeling at the clause-text level. We also found cross-domain duplicates: 49 groups totaling 405 clauses appear in both `national` and `local`. These can cause train–test leakage under naive random splits; therefore, our experiments incorporate duplicate-aware and domain-aware splitting strategies (Section 4).

Template reuse is more pronounced in `national` than in `local` (Table 4), suggesting that leakage control is especially important when models are trained primarily on national templates.

3.6 Corpus Statistics

To facilitate comparability with LEDGAR [17], we compute clause- and token-level statistics. For Japanese tokenization, we use Sudachi (Mode C) [15]. The clause bodies contain 1,869,567 tokens in total, with a vocabulary size of 4,105. The average number of tokens per clause is 201.4 (std. dev. 200.6), and the average number of clauses per document is 43.8 (std. dev. 16.4).

Table 5 compares these statistics with the cleaned LEDGAR corpus as reported in [17]. While JGovCCC-PDL is substantially smaller in raw scale, it provides a focused, redistributable Japanese resource with a controlled label inventory for clause-type classification.

4 Task Definition and Experimental Setup

4.1 Task Definition

In this study, we address *clause-level classification* (*clause-type classification*) for public procurement contracts, where the input is an individual clause and the goal is to predict its clause type. Concretely, we treat the fine-grained clause-type label (`lv2`; 50 classes) assigned to each clause in the corpus as the target variable, and formulate the task as a single-label multi-class classification problem. In addition, the corpus provides a corresponding coarse-grained category label (`lv1`; 9 classes) for each clause; however, unless otherwise noted, the experiments reported in the following sections focus primarily on the fine-grained labels (`lv2`).

4.2 Preprocessing and Input Representation

To prioritize reproducibility, we apply minimal text normalization (e.g., full-width to half-width spaces, whitespace collapsing, newline unification). We evaluate three input settings:

- **body**: Use only the clause body text.
- **heading**: Use only the clause heading.
- **heading+body**: Concatenate the heading and the body.

When using Transformer-based models, `heading+body` is not tokenized as a plain string concatenation; instead, we tokenize it as a text pair (heading, body) to avoid model-specific differences in special-token handling.

Table 3: Representative document and clause examples in JGovCCC-PDL. Clause content is abbreviated and translated into English by the authors.

Source document family	Example heading	Fine-grained label	Abbreviated clause content / function
Public works construction conditions	General provisions	General Provisions	The parties must perform the construction contract in accordance with the agreement and design documents; the contractor completes and delivers the work, and the ordering party pays the contract price.
Maintenance-service contract	Inspection	Inspection, Delivery, and Acceptance	After completing all or part of the work, the contractor submits a completion report; the ordering party inspects within a fixed period, and failed work must be corrected and re-inspected.
Public works construction conditions	Change of design documents	Contract Modification	The ordering party may notify the contractor of changes to design documents and, where necessary, adjust the term or price or bear necessary costs caused to the contractor.
Forestry/timber sale conditions	Forest conservation measures	Environmental and Waste	During cutting, removal, or related work, the contractor or buyer must prevent forest degradation and river pollution; the agency may require necessary measures at the counterparty's cost.
Research/content contract with Bayh-Dole clauses	Free implementation by government	Intellectual Property	The government or a designated third party may use intellectual-property rights in content created under the commissioned work when necessary to achieve the contract purpose.
System operation and maintenance contract	Access to ordering-party systems	Personal Information / Information Security	When accessing the ordering party's systems, the contractor must follow the ordering party's instructions about the information type, scope, and access method.

Table 4: Duplicate burden by domain based on exact matching of minimally normalized clause body text.

Domain	#Clauses	#Unique bodies	Dup. rate
`national`	6,577	3,659	44.4%
`local`	2,704	2,009	25.7%
All	9,281	5,619	39.5%

Table 5: Corpus-scale comparison between JGovCCC-PDL and the cleaned LEDGAR corpus.

Statistic	JGovCCC-PDL	LEDGAR (clean)
No. of documents/contracts	212	60,540
No. of clause/provision instances	9,281	846,274
No. of labels	50	12,608
No. of tokens	1,869,567	104,990,418
Vocabulary size	4,105	52,098
Avg. tokens per instance (std. dev.)	201.4 (200.6)	124 (104)
Avg. instances per doc. (std. dev.)	43.8 (16.4)	13 (20)

LEDGAR is a multi-label corpus, while JGovCCC-PDL assigns a single fine-grained label (1v2) per clause. Token statistics are shown as reported in each corpus and are not strictly comparable across languages and tokenization.

4.3 Handling Duplicates and Label Conflicts

Public procurement templates yield many repeated clauses. We identify duplicates by exact matching of minimally normalized clause body text. We define *conflict duplicates* as identical clause bodies assigned different labels and exclude them; none were observed in our experimental subset. To reduce train–test leakage from boilerplate reuse, we additionally report settings that remove exact *(body, label)* duplicates.

In this paper, *leakage-resistant evaluation* refers to two concrete design choices motivated by the template-heavy nature of public procurement contracts: (i) duplicate-aware evaluation, where identical clause bodies are not allowed to trivially overlap between training and test data or are removed as exact *(body, label)* duplicates, and (ii) domain-transfer evaluation between the `national` and `local` subsets. These choices matter because 39.5% of clauses are reoccurrences of identical clause bodies, and 49 duplicate groups (405 clauses) are shared across the two domains. Under naive random splits, such overlap can overestimate real generalization.

4.4 Splitting Strategies

Following LEDGAR, we use a 70%/10%/20% train/dev/test split as the default. In this study, we mainly report the following two families of evaluation settings:

- **Random split (random)**: a random split with label stratification.
- **Domain transfer (transfer)**: train on `national` and test on `local` (`national` → `local`), and vice versa (`local` → `national`).

The transfer experiments are *open-set*: we do not remove test instances whose labels are absent from the training domain. This asymmetry mainly affects `local` → `national`; after deduplication, three labels (*Definitions*, *Contract Documents / Specifications*, and *Disclosure / Publication*) appear only in the `national` subset.

4.5 Compared Methods

We compare the following classifiers:

- **Majority**: always predicts the most frequent label in the training data.

- **Label-name match**: predicts the first label whose name appears in the input text (preferring longer names); otherwise falls back to **Majority**.
- **TF-IDF + Logistic Regression (LogReg)**: character n-gram TF-IDF ($n = 3$–5) with multinomial logistic regression (`saga`). We tune the regularization parameter C on the dev set to maximize macro-F1.
- **Transformer fine-tuning**: we fine-tune a Japanese ModernBERT model (`sbintuitions/modernbert-ja-130m`) [16, 20] end-to-end for clause classification. For pooling, we use the mean of the last four layers (`last4mean`). The Transformer input length is capped at 256 tokens, so longer clauses are truncated.

For LogReg, the dev set selects $C \in \{1.0, 2.0\}$ by macro-F1. For ModernBERT, we fine-tune for 3 epochs with learning rate 2×10^{-5}, weight decay 0.01, training batch size 2, evaluation batch size 16, and checkpoint selection by dev macro-F1. The truncation setting above is important because JGovCCC-PDL clauses are often long (average 201.4 Sudachi tokens per clause; 95th percentile 629).

4.6 Evaluation Metrics

We report Accuracy as well as macro-averaged Precision/Recall/F1 to account for label imbalance. We also report micro-averaged Precision/Recall/F1 for completeness. Because our setting is single-label multi-class classification, micro-F1 coincides with Accuracy. Unless otherwise specified, we use micro-F1 (i.e., Accuracy) as the primary metric in this paper.

5 Results

5.1 Classification under Random Splits

Table 6 reports the results under random splits. When using only the clause body as input, Majority and Label-name match remain low (around 0.07–0.09 micro-F1), whereas LogReg achieves a high micro-F1 over 0.96. Furthermore, the Transformer-based model (ModernBERT) consistently outperforms LogReg in terms of micro-F1.

5.2 Effect of Duplicate Removal

After removing exact *(body, label)* duplicates, the micro-F1 of LogReg decreases from 0.9688 to 0.9200, and the micro-F1 of ModernBERT drops from 0.9790 to 0.9520 (Table 6). This gap suggests that the reuse of boilerplate clauses in public procurement contracts can induce train–test leakage under random splits. At the same time, performance remains around 0.90 even after duplicate removal, indicating that clause types in this corpus are, to a certain extent, distinguishable from lexical and stylistic cues.

5.3 Impact of Heading Information

Including headings in the input (`heading+body`) improves performance, most notably for Label-name match (0.0756 → 0.1068). This is consistent with the intuition that headings are more likely than clause bodies to contain expressions closely related to clause types. We also observe improvements for LogReg (0.9200 → 0.9342) and for ModernBERT (0.9520 → 0.9671), suggesting that headings provide useful complementary signals to the clause body.

5.4 Domain Transfer

Table 6 presents transfer evaluation results between `national` and `local`. For both LogReg and ModernBERT, performance drops substantially compared to random splits. These transfer results are open-set: in the `local` → `national` direction, 24 deduplicated test clauses are assigned to three labels that never appear in the local training data. This indicates that differences in clause expressions and clause organization across institution types (national government vs. local municipalities) make generalization in clause-type classification more challenging.

To make the treatment of label imbalance more explicit, Table 7 reports Accuracy together with macro-averaged Precision/Recall/F1 for the two strongest models on the main settings. The overall ranking is unchanged, but macro-F1 drops more sharply under deduplication and transfer than accuracy alone would suggest.

6 Discussion

6.1 Duplicates and Evaluation Leakage

JGovCCC-PDL contains many boilerplate clauses reused across documents and institutions: 39.5% of clauses are reoccurrences of identical clause bodies, and 49 duplicate groups (405 clauses) span both `national` and `local`. As a result, naive random splits can place identical text in both training and test data. The performance drop after duplicate removal, therefore, indicates that such overlap can inflate evaluation scores, while the further drop under `national` ↔ `local` transfer shows that lexical overlap is not the same as robust generalization. For more realistic estimates of generalization, we recommend (i) duplicate-aware evaluation and (ii) leakage-resistant splits at the document, institution, and duplicate-group levels.

6.2 Role of Headings

Adding headings improves performance across models, and the gain for the string-matching baseline indicates that headings often state clause functions explicitly. This aligns with practical use: headings are lightweight metadata that help organize and retrieve clauses. However, headings are not always available and their phrasing varies across templates, so heading-heavy models may be less robust; reporting heading-only, body-only, and combined inputs remains important depending on the intended deployment setting.

6.3 Domain Differences

Transfer results reveal substantial performance degradation between `national` and `local`, indicating domain shift across institution types. Such differences likely reflect institution-specific terminology, referenced rules/forms, and conventions in clause ordering and granularity. The `local` → `national` direction is slightly harder. After deduplication, three fine-grained labels (*Definitions*, *Contract Documents / Specifications*, and *Disclosure / Publication*) appear only in the `national` subset, contributing 24 test clauses in the open-set `local` → `national` evaluation. This asymmetry partly explains the directional gap, in addition to broader domain differences. These results motivate the use of transfer-style evaluation as a stress test and the exploration of domain adaptation and broader institutional coverage.

Table 6: Clause-type classification results (Accuracy (micro-F1)).

Setting	Majority	Label-name	LogReg	ModernBERT
random (body, with duplicates)	0.0727	0.0813	**0.9688**	**0.9790**
random (body, duplicates removed)	0.0854	0.0756	0.9200	0.9520
random (heading+body, duplicates removed)	0.0854	**0.1068**	0.9342	0.9671
transfer `national`→`local` (heading+body, duplicates removed)	**0.0858**	0.0725	0.7463	0.8058
transfer `local`→`national` (heading+body, duplicates removed)	0.0674	0.0851	0.6971	0.7754

Table 7: Accuracy and macro-averaged Precision/Recall/F1 for the main settings.

	LogReg				ModernBERT			
Setting	Acc.	P	R	F1	Acc.	P	R	F1
random (body, with duplicates)	**0.9688**	**0.9699**	**0.9496**	**0.9584**	**0.9790**	**0.9719**	**0.9707**	**0.9705**
random (heading+body, duplicates removed)	0.9342	0.9577	0.8785	0.9053	0.9671	0.9536	0.9425	0.9439
transfer `national`→`local` (heading+body, duplicates removed)	0.7463	0.7468	0.6587	0.6632	0.8058	0.7050	0.7307	0.6940
transfer `local`→`national` (heading+body, duplicates removed)	0.6971	0.6829	0.5734	0.5964	0.7754	0.6721	0.6714	0.6601

6.4 Scope and Representativeness

The corpus should be interpreted as a licensed, template-focused benchmark rather than as a statistically representative sample of all Japanese contracts. The 212 documents expose recurring clause functions across 32 institutions and support a reproducible 50-class classification task, but they do not justify broad claims about clause-type prevalence in Japanese private contracting or all public procurement practice. This scope choice was deliberate: project-specific executed contracts may contain richer factual details, but they also create greater licensing, privacy, and redaction risks. Future expansion should prioritize additional local governments, procurement families, temporal coverage, and, where licensing and privacy controls permit, project-specific contract records.

6.5 Brief Error Analysis

A brief manual inspection of misclassified clauses suggests that remaining errors are often semantically local rather than arbitrary. First, some clauses bundle multiple legal effects in a single article, especially remedial clauses that combine damages language with explicit penalty provisions. Second, several confusions arise between neighboring operational labels because the same clause simultaneously describes receipt, conformity checking, and custody of supplied items. Third, confidentiality clauses often explicitly enumerate personal data and protected information, which makes the lexical boundary between confidentiality and information-security labels less clear than the functional distinction.

These examples suggest that future gains may come from incorporating wider document context and from exploiting the label hierarchy during prediction.

6.6 Comparison of Classical Models and Transformers

LogReg based on character *n*-gram TF-IDF achieves strong micro-F1 even after duplicate removal. Because public procurement contracts contain many formulaic phrases and fixed expressions, character *n*-grams, without relying on morphological analysis, can directly capture informative keywords even under orthographic variation (e.g., *termination*, *damages*, *antisocial forces*).

Meanwhile, the Transformer model (ModernBERT) consistently outperforms LogReg, with particularly large gains under (i) the setting with heading information and (ii) the domain transfer settings. This suggests that ModernBERT may be relatively more robust to contextual cues and paraphrases even when expressions and vocabulary differ across domains. However, both models degrade substantially under transfer compared to random splits, indicating that domain shift driven by institutional and conventional differences between `national` and `local` remains the primary bottleneck.

7 Conclusion

We introduced JGovCCC-PDL, a redistributable Japanese clause corpus for public procurement contracts. JGovCCC-PDL is built from 212 publicly available standard documents collected from 32 public institutions and segmented into 9,281 clauses, each annotated with a two-level label hierarchy (50 fine-grained clause types grouped into 9 coarse categories).

Baseline experiments show that character *n*-gram TF-IDF with logistic regression achieves strong performance without heavy Japanese-specific preprocessing, while a Transformer model (ModernBERT) provides further gains, especially when incorporating headings and under domain transfer. At the same time, performance drops after duplicate removal and under transfer settings, highlighting the importance of leakage-resistant evaluation and domain robustness. We emphasize that the primary contribution of this paper is the redistributable resource and benchmark rather than a novel model architecture. Because clause-type classification is a prerequisite for downstream contract review, clause retrieval, issue extraction, and later reasoning, we expect JGovCCC-PDL to serve as a foundation for Japanese contract-oriented Legal NLP. Future work includes expanding institutional coverage and investigating domain adaptation and hierarchical learning methods to better handle cross-institution variation.

Table 8: Representative error patterns in clause-type classification.

Gold label	Predicted label	Typical reason
Damages and Liability	Penalties	one clause combines compensatory damages language with a fixed penalty or liquidated-damages provision
Inspection, Delivery, and Acceptance	Materials, Goods, and Loaned Items	handover clauses often mix receipt, conformity checking, and custody of supplied items
Confidentiality	Personal Information / Information Security	confidentiality clauses explicitly enumerate personal data and protected information, creating strong lexical overlap

References

[1] Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. *ACM Comput. Surv.* 58, 6, Article 163 (Dec. 2025), 37 pages. doi:10.1145/3777009

[2] Zhihan Cao, Fumihito Nishino, Hiroaki Yamada, Ha Thanh Nguyen, Yusuke Miyao, and Ken Satoh. 2025. JBE-QA: Japanese Bar Exam QA Dataset for Assessing Legal Domain Knowledge. *arXiv* abs/2511.22869 (2025). https://arxiv.org/abs/2511.22869

[3] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. 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)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4310–4330. doi:10.18653/v1/2022.acl-long.297

[4] Digital Agency, Government of Japan. 2024. Public Data License (Version 1.0). https://www.digital.go.jp/en/resources/open_data/public_data_license_v1.0. Accessed: 2026-01-31.

[5] Ruka Funaki, Yusuke Nagata, Kohei Suenaga, and Shinsuke Mori. 2020. A Contract Corpus for Recognizing Rights and Obligations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2045–2053. https://aclanthology.org/2020.lrec-1.251/

[6] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *NeurIPS* (2021).

[7] Mi-Young Kim, Juliano Rabelo, Housam Khalifa Bashier Babiker, Md Abed Rahman, and Randy Goebel. 2024. Legal Information Retrieval and Entailment Using Transformer-based Approaches. *The Review of Socionetwork Strategies* 18, 1 (April 2024), 101–121. doi:10.1007/s12626-023-00153-z

[8] Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 1907–1919. doi:10.18653/v1/2021.findings-emnlp.164

[9] Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service. *Artificial Intelligence and Law* 27 (2019), 117–139. doi:10.1007/s10506-019-09243-2

[10] Keisuke Miyazaki, Hiroaki Yamada, and Takenobu Tokunaga. 2022. Cross-domain Analysis on Japanese Legal Pretrained Language Models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*. Association for Computational Linguistics, 274–281. https://aclanthology.org/2022.findings-aacl.26/

[11] Damith Premasiri, Tharindu Ranasinghe, Ruslan Mitkov, Mo El-Haj, and Ingo Frommholz. 2025. Survey on legal information extraction: current status and open challenges. *Knowledge and Information Systems* 67, 12 (Dec. 2025), 11287–11358. doi:10.1007/s10115-025-02600-5

[12] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. *The Review of Socionetwork Strategies* 16 (2022), 111–133. doi:10.1007/s12626-022-00105-z

[13] Abhilasha Sancheti, Aparna Garimella, Balaji Vasan Srinivasan, and Rachel Rudinger. 2022. Agent-Specific Deontic Modality Detection in Legal Language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11563–11579. doi:10.18653/v1/2022.emnlp-main.795

[14] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 1048–1064. doi:10.18653/v1/2022.aacl-main.77

[15] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: a Japanese Tokenizer for Business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan, 7-12), Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Paris, France.

[16] Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. 2025. ModernBERT-Ja. https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a

[17] Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1235–1241. https://aclanthology.org/2020.lrec-1.155/

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[19] Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 16369–16382. doi:10.18653/v1/2023.emnlp-main.1019

[20] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 2526–2547. doi:10.18653/v1/2025.acl-long.127

[21] Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. *Artificial Intelligence and Law* 27 (2019), 141–170. doi:10.1007/s10506-019-09242-3

[22] Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Akira Tokutsu, Keisuke Takeshita, and Mihoko Sumida. 2025. Japanese tort-case dataset for rationale-supported legal judgment prediction. *Artificial Intelligence and Law* 33 (2025), 783–807. doi:10.1007/s10506-024-09402-0