# Formalizing FOIA Exemption Standards as Computable Legal Norms: A Framework for LLM-Assisted Sensitivity Review

Romana Afroze
Delaware Law School
Wilmington, Delaware, USA
rafroze@widener.edu

## ABSTRACT

Governments worldwide are deploying artificial intelligence to assist with Freedom of Information Act (FOIA) and Access to Information (ATI) review, yet existing systems treat sensitivity classification as a pure machine-learning task, ignoring the normative legal structure encoded in the exemption doctrine. This paper argues that FOIA exemptions are not labels but *computable legal norms*: structured balancing tests whose elements can be formalized using deontic logic and ontological modelling. This paper proposes a two-layer framework—a Legal Norm Ontology (LNO) for the nine FOIA exemptions and a Norm-Informed Prompting Architecture (NIPA) that operationalizes the ontology within large language model (LLM) pipelines. NIPA is evaluated against two baselines on a curated corpus of 847 judicially reviewed FOIA redaction decisions drawn from D.C. Circuit litigation records, demonstrating a 14.4-percentage-point improvement in exemption classification precision over the fine-tuned baseline, and a 22.1-percentage-point gain in legal reasoning alignment (LRA—the proportion of criterion-level assessments whose rationale matches the ground-truth annotation protocol) over the zero-shot baseline. A Disclosure-of-Reasoning (DoR) standard is derived from administrative due process doctrine and argued to constitute a *necessary condition* for lawful deployment—not merely a governance aspiration. These findings have direct implications for AI governance frameworks in open government contexts globally.

## KEYWORDS

FOIA; sensitivity review; computable legal norms; deontic logic; large language models; explainability; open government; access to information

**CCS Concepts:** • Computing methodologies: Natural language processing; • Applied computing: Law, social and behavioral sciences; • Theory of computation: Logic.

## 1 INTRODUCTION

The Freedom of Information Act, 5 U.S.C. § 552, and its international counterparts—Canada's *Access to Information Act*, the United Kingdom's *Freedom of Information Act 2000*, and analogous statutes across more than 130 jurisdictions—represent among the most consequential transparency mechanisms in democratic governance. They establish a presumption in favour of disclosure, subject to a defined set of exemptions that permit agencies to withhold information whose release would cause specific categories of harm. In the United States alone, federal agencies received approximately 1.2 million FOIA requests in fiscal year 2023, processing more than 1.1 million requests including backlog [19].

The sheer scale of this enterprise has driven sustained investment in technology-assisted review. Agencies including the Department of Justice, the Department of State, and the National Archives and Records Administration have piloted or deployed AI tools for document triage, deduplication, metadata enrichment, and—most consequentially—automated sensitivity review and redaction recommendation.

Yet existing deployments share a critical architectural flaw: they treat FOIA sensitivity review as a machine-learning classification problem. This framing, while computationally tractable, is legally inadequate. FOIA exemptions are not discrete categorical labels; they are normative legal standards encoding multi-factor balancing tests, statutory terms of art, judicially elaborated doctrines, and agency-specific implementing guidance. Exemption 6, for instance, requires an assessment of whether disclosure would constitute a "clearly unwarranted" invasion of personal privacy, a phrase elaborated across hundreds of judicial decisions that cannot be reduced to a binary signal without loss of legally material information [4].

This paper makes three principal contributions. First, this paper argues that the correct conceptual frame for AI-assisted FOIA review is not classification but *norm operationalization*: the task is to instantiate computable representations of legal norms and reason over documents with respect to those norms. Second, this paper develops a Legal Norm Ontology (LNO) for the nine FOIA exemptions, drawing on deontic logic, statutory interpretation theory, and judicial precedent, alongside a Norm-Informed Prompting Architecture (NIPA) that structures LLM inference around the ontology. Third, this paper derives a Disclosure-of-Reasoning (DoR) standard from administrative due process doctrine—a legally grounded requirement that AI-assisted redaction systems generate human-auditable explanations for each withholding recommendation—and argues this standard constitutes a necessary condition for lawful deployment.

## 2 BACKGROUND

### 2.1 FOIA Exemption Doctrine

The Freedom of Information Act establishes nine categories of exemption to the general presumption of disclosure [2,


*ICAIL '26, June 2026, Singapore*

ACM ISBN 978-1-4503-XXXX-X/26/06
https://doi.org/10.1145/XXXXXXX.XXXXXXX

5 U.S.C. § 552(b)]. These exemptions are not self-executing; their application requires the exercise of legal judgment across multiple dimensions. Three structural features are legally significant for AI design.

First, exemptions are *overlapping and cumulative*: a single passage may implicate multiple exemptions, and agencies frequently invoke multiple exemptions as independent and alternative grounds. Second, exemptions are scope-limited by the principle of *segregability*: agencies must disclose all reasonably segregable non-exempt information, requiring assessment at sub-document granularity [2]. Third, the application of most exemptions is *foreseeable harm-sensitive*: the FOIA Improvement Act of 2016 amended the Act to require agencies to withhold only where they "reasonably foresee that disclosure would harm an interest protected by an exemption" or disclosure is prohibited by law—a standard that introduces a causal harm-prediction element that pure classification models cannot represent.

Judicial review of FOIA withholdings proceeds de novo, with the burden on the agency to sustain its exemption claims [2]. This means that AI-generated withholding recommendations must ultimately be capable of legal justification in litigation.

## 2.2 AI in Sensitivity Review: Prior Work

Technology-assisted review (TAR) methods from e-discovery have been adapted for FOIA review, primarily through predictive coding approaches that use relevance feedback to train binary classifiers on small labelled seed sets [10]. McDonald et al. [13, 14] have produced the most systematic body of work on automated sensitivity review, developing datasets from UK National Archives materials and evaluating a range of classifier architectures. Their findings demonstrate that fine-tuned transformer models can achieve competitive sensitivity detection performance but note persistent difficulties with context-dependent sensitivity—precisely the cases requiring legal norm reasoning.

Recent evaluations of GPT-4 reflect this architectural dichotomy. On categorical named-entity de-identification tasks in the medical domain—which share structural features with Exemption 6 PII redaction but differ substantially in the normative complexity required—zero-shot prompting achieves near-ceiling precision (0.99 in clinical-note de-identification against labelled PII categories [7, 12]). Those figures are not directly comparable to FOIA performance: the medical tasks require identifying *what* is sensitive within a well-defined taxonomy, whereas FOIA requires also determining *whether and why* withholding is legally justified under a multi-factor balancing test. Accordingly, the baseline GPT-ZS precision of 0.721 observed in this study should be understood against the specific difficulty of normative FOIA reasoning, not against medical de-identification benchmarks. Separately, Dahl et al. [8] document substantial hallucination rates in case-law factual question answering—a different failure mode from normative balancing errors, but one that similarly cautions against treating LLM outputs as authoritative in legal contexts without structured scaffolding. To the best of the author's knowledge, no prior work has grounded LLM-based FOIA review in a formal ontological representation of exemption doctrine.

## 2.3 Computable Legal Norms

The formalization of legal norms has a substantial scholarly lineage. Susskind [18] proposed expert systems for legal advice; Sergot et al. [17] formalized the British Nationality Act in Prolog. The contemporary literature distinguishes between rule-based and standard-based legal norms: rules specify conditions and consequences with relative determinacy, while standards require contextual judgment [11]. Exemption doctrine sits primarily in the standards category, which has historically resisted clean formalization.

Deontic logic—the logic of obligation, permission, and prohibition [20]—provides the canonical formal language for expressing normative content. Work in normative multi-agent systems has extended deontic logic to support conditional obligations, contrary-to-duty structures, and defeasibility [9], all of which are present in FOIA doctrine. Robaldo et al. [16] and Palmirani et al. [15] have developed ontologies for legal norms within the Akoma Ntoso and LKIF frameworks. The LNO framework builds on this tradition while addressing the specific structural features of FOIA exemption doctrine identified above.

# 3 A LEGAL NORM ONTOLOGY FOR FOIA EXEMPTIONS

## 3.1 Conceptual Foundation

A Legal Norm Ontology (LNO) for FOIA is defined as a five-tuple $\langle E, C, R, B, H \rangle$, where: $E = \{e_1, \ldots, e_9\}$ is the set of FOIA exemptions; $C$ is a set of exemption criteria, where each criterion $c \in C$ is a tuple $\langle$predicate, threshold, modality$\rangle$; $R \subseteq E \times C$ is the mapping from exemptions to their constituent criteria; $B \subseteq C \times C$ is the balancing relation, specifying criteria that stand in proportionality tension; and $H$ is the harm-foreseeability predicate introduced by the 2016 FOIA Improvement Act.

Three deontic modalities are distinguished: **OBL** (the agency is obligated to assess this criterion), **PERM** (the agency is permitted to invoke this criterion), and **PROHIB** (the agency is prohibited from withholding absent this criterion). For Exemption 6, for example, the privacy-interest criterion is **OBL**; the public-interest criterion is also **OBL** by virtue of the balancing requirement. Balancing relation $B$ maps these two criteria against each other, encoding the proportionality structure.

## 3.2 Formalizing the Nine Exemptions

The following list presents the full LNO specification for all nine FOIA exemptions, grounded in 5 U.S.C. § 552(b)(1)–(9) [2], the FOIA Improvement Act of 2016 (harm-foreseeability predicate $H$), and *DOJ v. Reporters Committee*, 489 U.S. 749 (1989) [4] (Exemptions 6 and 7(C) balancing thresholds). Three illustrative cases demonstrate the expressiveness of the framework.

- **Exemption 1 (Classified Information):** Requires EO classification category membership (**OBL**) and disclosure reasonably expected to damage national security (**OBL**). Conjunctive relation; harm-foreseeability ($H$) applies.
- **Exemption 2 (Internal Personnel Rules):** Information relates solely to internal personnel rules or practices (**OBL**). $H$ applies.

- **Exemption 3 (Statutory Prohibition):** Disclosure prohibited by statute specifying criteria for withholding or particular matters to be withheld (**PROHIB**). Per se prohibition; $H$ does not apply.
- **Exemption 4 (Trade Secrets & Commercial Information):** Constitutes a trade secret, or privileged or confidential commercial or financial information obtained from a person (**OBL**). $H$ applies.
- **Exemption 5 (Deliberative Process Privilege):** Requires (i) pre-decisional AND deliberative [DP sub-test]; OR (ii) confidential attorney-client communication for legal advice [AC sub-test]; OR (iii) work product prepared in anticipation of litigation [WP sub-test]. Disjunctive inference (**OBL** per sub-test); any one sub-graph sufficient; $H$ applies.
- **Exemption 6 (Personal Privacy):** Requires an identified personal privacy interest (**OBL**) and an assessed public interest in governmental conduct (**OBL**). Asymmetric default to privacy; "*clearly* unwarranted" invasion threshold; $B(c_{\text{priv}}, c_{\text{pub}})$; $H$ applies.
- **Exemption 7 (Law Enforcement Records):** Records compiled for law enforcement purposes (**OBL**), with sub-criteria 7(A)–(F) covering interference, fair trial, unwarranted privacy invasion, confidential sources, techniques, and endangerment. Balancing for 7(C) applies an "unwarranted" threshold—lower than Exemption 6's "clearly unwarranted" standard; $H$ applies.
- **Exemption 8 (Financial Institution Records):** Contained in or related to examination, operating, or condition reports of financial institutions prepared by or for a supervisory agency (**OBL**). $H$ applies.
- **Exemption 9 (Geological Information):** Geological and geophysical information and data, including maps, concerning wells (**OBL**). $H$ applies.

### 3.3 Segregability and Sub-Document Granularity

The segregability requirement imposes a structurally distinct challenge: the LNO must be applicable at the passage level. This is operationalized by defining a segmentation function $\sigma$ that partitions documents into *candidate-withholding units* (CWUs)—contiguous text segments at the paragraph level, falling back to the sentence level where paragraph boundaries are absent or ambiguous. The same $\sigma$ was applied uniformly to all three configurations: BERT-FT receives individual CWUs as classification inputs, and GPT-ZS and NIPA receive them as the unit of prompting. This ensures that precision comparisons across configurations reflect model behaviour rather than segmentation differences. A per-document CWU count breakdown is not reported here; this is acknowledged as a reproducibility limitation, and future work should publish the full segmentation log alongside the corpus release. The segregability obligation is modelled as a constraint: for each CWU, the LNO is applied independently, and the withholding recommendation is narrowed to the minimum set of CWUs satisfying the applicable exemption criteria.

## 4 NORM-INFORMED PROMPTING ARCHITECTURE (NIPA)

### 4.1 Architectural Design

The Norm-Informed Prompting Architecture operationalizes the LNO within an LLM inference pipeline through a structured three-stage process. Stage 1 (*Segmentation*) applies $\sigma$ to produce a set of candidate passages. Stage 2 (*Criterion-Conditioned Assessment*) generates, for each CWU and each potentially applicable exemption, a structured prompt derived directly from the LNO—instantiating the criteria, modalities, and balancing relations as explicit reasoning instructions. Stage 3 (*Norm Aggregation*) applies the inference rules encoded in the LNO to aggregate criterion-level assessments into an exemption-level withholding recommendation with supporting rationale.

The key design principle distinguishing NIPA from baseline prompt engineering is *norm explicitness*: rather than prompting the LLM with a free-text description of the exemption, NIPA decomposes the exemption into its constituent criteria and prompts for each criterion assessment separately, then uses the LNO's inference rules to combine assessments. This approach operationalizes recent findings in chain-of-thought prompting [22] within a legally structured framework.

For the Exemption 6 balancing test, for example, the NIPA Stage 2 prompt specifies: (a) the identity of the individual(s) whose privacy interest is at stake; (b) the nature of the information and its sensitivity category; (c) the counterfactual question of what specific public interest in governmental conduct would be served by disclosure; and (d) the asymmetric default in favour of privacy protection. The model assesses each element separately before generating a recommendation. This structure is generated programmatically from the LNO, not hand-crafted per exemption.

### 4.2 Corpus, Train/Test Split, and Evaluation Design

A new evaluation corpus was constructed comprising 847 FOIA withholdings reviewed by courts. The dataset was assembled through three steps: (i) collection of Vaughn index submissions filed in FOIA litigation between 2015 and 2024 before the D.C. Circuit and D.C. District Court, sourced from PACER and the National Security Archive's FOIA litigation database; (ii) transformation of redactions into reconstructed withheld text segments; and (iii) incorporation of findings from Inspector General reports evaluating agency FOIA processing practices.

*Train/test partition.* The corpus was partitioned into a training set of 677 records (80%) and a held-out test set of 170 records (20%) using stratified sampling to preserve the exemption-type distribution across both splits. To guard against data leakage arising from chronologically contiguous litigation, the partition enforces a *temporal boundary*: all records from FOIA cases whose earliest docketed filing pre-dates January 1, 2022 are eligible for the training set; records from cases filed from that date onward are assigned exclusively to the test set. Because Vaughn index entries in the same case share procedural context, this boundary ensures that no test-set CWU is drawn from a case whose companion records appear in the training set. The BERT-FT baseline was fine-tuned

*exclusively* on the 677 training-set records; all LLM-based configurations (GPT-ZS and NIPA) were evaluated *zero-shot* on the 170 test-set records, receiving no exposure to training-set labels. This design ensures ecological validity: the BERT-FT advantage of in-domain fine-tuning is preserved, while the LLM configurations are assessed under conditions that reflect the zero-shot deployment scenario most relevant to agency practice, where labelled corpora are rarely available at the point of system procurement.

Ground-truth labels were assigned at the CWU level by three legal reviewers: one attorney with FOIA litigation experience and two attorneys specializing in FOIA review. Annotation followed a structured protocol grounded in LNO criteria. Inter-annotator agreement was measured using Cohen's $\kappa$, yielding $\kappa$=0.76 for exemption-level labels and $\kappa$=0.68 for criterion-level assessments, both reflecting substantial agreement. It is acknowledged that $\kappa$=0.68 is at the lower end of the "substantial" range, and that a per-exemption $\kappa$ breakdown— particularly for the Exemption 5 and 6 sub-tests where criterion-level judgment is most contested—would allow more precise calibration of annotation reliability. This breakdown is not available for the current study and is flagged as a limitation; future corpus releases should include per-exemption inter-annotator statistics. Disagreements were resolved by majority determination.

The NIPA framework was evaluated against two baselines: (i) a BERT-base classifier fine-tuned on the training partition (BERT-FT); and (ii) a zero-shot GPT-4o configuration employing free-text exemption descriptions (GPT-ZS). All LLM experiments used Claude 3.5 Sonnet as the underlying model—selected for its strong adherence to structured prompt instructions—at temperature 0 to ensure reproducibility.

## 4.3 Results

The following results are reported as macro-averaged precision and Legal Reasoning Alignment (LRA) on the held-out test set ($n$=170). LRA measures criterion-level rationale agreement with the annotation protocol. LRA is not applicable (N/A) to BERT-FT, which produces a classification label only and generates no criterion-level rationale.

- **BERT-FT:** Macro Precision: 0.703; LRA: N/A.
- **GPT-ZS (baseline):** Macro Precision: 0.721; LRA: baseline.
- **NIPA (this work):** Macro Precision: 0.847; LRA: +22.1 pp over GPT-ZS baseline.

NIPA achieves a macro-average precision of 0.847—a 14.4-percentage-point improvement over the fine-tuned BERT-FT baseline (0.703)—and outperforms the GPT-ZS baseline (0.721) by 12.6 percentage points, despite receiving no in-domain supervision. The most significant performance differences appear in Exemptions 5 and 6, which call for multi-factor balancing and contextual judgment. The gap narrows on Exemption 1 (a more categorical determination), consistent with the prediction that norm-informed prompting provides the greatest benefit where the legal structure is most complex.

Error analysis revealed two principal failure modes. First, NIPA underperforms on passages where the withholding depends on classified contextual information not present in the document itself—a structural limitation of text-only models that no prompting approach can overcome. Second, NIPA occasionally conflates the deliberative process and attorney-client sub-tests within Exemption 5. The source of this failure is architectural: although the LNO encodes the three sub-tests (DP, AC, WP) as formally separate sub-graphs connected by a disjunctive inference rule, the Stage 2 prompt presents all three sub-tests to the model within a single context window, and the model's attention mechanism does not enforce the sub-graph boundaries. The result is that passages satisfying only the AC sub-test are occasionally labelled as satisfying the DP sub-test as well, inflating the criterion-level assessment record in a way that would not survive de novo judicial review. The remedy is to issue three sequential Stage 2 calls for Exemption 5—one per sub-test—and apply the disjunctive inference rule at Stage 3, rather than presenting all three within a single prompt. This architectural refinement is reserved for a future iteration of the system.

## 4.4 Ablation: Separating Ontology from Task Decomposition

Reviewers correctly observed that NIPA bundles two distinct innovations: (i) formal ontological grounding via the LNO, and (ii) criterion-level task decomposition in the prompting architecture. To isolate each contribution, two ablation conditions are defined alongside the full system: **NIPA-Onto** uses the LNO's formal structure to inform a single holistic exemption-level prompt (ontology without decomposition); **NIPA-Decomp** decomposes assessment into per-criterion prompts using free-text statutory descriptions, without instantiating the formal LNO modalities or balancing relations (decomposition without ontology); **NIPA-Full** combines both components (the system whose results are reported in Section 4.3 above).

Across Exemptions 5, 6, and 7(C)—the three exemptions requiring multi-factor balancing—NIPA-Decomp outperforms NIPA-Onto on exemption-level precision, indicating that explicit criterion separation is the primary driver of the performance gain over the GPT-ZS baseline. However, NIPA-Full consistently exceeds NIPA-Decomp on LRA scores, confirming that formal ontological grounding adds measurable reasoning structure beyond free-text decomposition alone. This contribution is most pronounced for Exemption 5, where the LNO's disjunctive sub-graph structure correctly constrains the model to distinguish DP, AC, and WP sub-tests that narrative decomposition tends to conflate.

*Audit trail implications.* The ablation has a direct bearing on DoR compliance that goes beyond performance metrics. NIPA-Onto, lacking criterion-level decomposition, produces a single holistic rationale that cannot be mapped back to individual exemption criteria—it therefore fails the first element of the DoR standard (the criterion-level assessment record) regardless of its exemption-level accuracy. NIPA-Decomp generates criterion-level outputs but without the formal modality and balancing structure of the LNO; its rationales are therefore not systematically auditable against the legal standard, producing outputs that may satisfy a superficial explainability check but cannot sustain de novo judicial review. Only NIPA-Full generates rationales that are both criterion-decomposed and formally

grounded in the LNO's modality and balancing structure—the combination necessary to produce a legally adequate criterion-level assessment record under the DoR standard. This result demonstrates that the two components of NIPA are not merely jointly beneficial for accuracy but are jointly *necessary* for legal compliance.

# 5 THE DISCLOSURE-OF-REASONING STANDARD

## 5.1 Due Process Foundations

The deployment of AI systems in FOIA review is not merely a technical question; it is a legal one. Administrative agencies making withholding decisions exercise public power with legal consequences for requestors: wrongful withholding denies the requestor information to which they are legally entitled and impairs the democratic functions FOIA is designed to serve. The Administrative Procedure Act establishes a baseline requiring agencies to provide reasoned explanations; the "arbitrary and capricious" standard under 5 U.S.C. § 706(2)(A) requires that an agency "examine the relevant data and articulate a satisfactory explanation for its action" [3].

In the FOIA context, this duty is heightened. Because courts review FOIA withholdings *de novo* [2], an agency cannot simply demonstrate that its automated process was "not arbitrary"; it must affirmatively prove the independent legal merit of every redacted passage. This evidentiary burden does not disappear because the initial withholding recommendation was generated by an AI system. Critically, the burden must be discharged contemporaneously: the APA's prohibition on post-hoc rationalization—established in *SEC v. Chenery Corp.*, 332 U.S. 194 (1947) [1], which held that an agency's action must be sustained, if at all, on the reasons it actually invoked at the time of decision—bars an agency from reconstructing a criterion-level justification after the fact to cover an AI recommendation that was generated without one. This is the inferential link from the *Motor Vehicle Mfrs.* "reasoned explanation" requirement to the DoR standard: a system that produces only a binary withholding label cannot supply the contemporaneous criterion-level record that *Chenery* requires; a system that generates criterion-level rationale at the moment of classification can.

AI-assisted redaction decisions trigger a Disclosure-of-Reasoning (DoR) obligation with three constituent elements. First, the agency must identify, for each withheld passage, which specific exemption criteria were assessed and how they were resolved—the *criterion-level assessment record*. Second, the agency must identify the AI system, model version, and prompting configuration used—the *system identification record*. Third, the agency must document the human review process through which the AI recommendation was accepted, modified, or rejected—the *oversight record*. This third element ensures that AI recommendations do not function as de facto final decisions, which would violate both the agency's statutory duty and the due process requirement that consequential decisions be made by accountable officials.

## 5.2 Implications for Governance

The DoR standard has practical implications for AI system design directly connected to the LNO framework. A system that generates criterion-level assessments—as NIPA does—is architecturally compatible with DoR compliance; the criterion-level output is precisely the information required for the criterion assessment record. A system that generates only a binary withholding recommendation is architecturally *incompatible* with DoR compliance, regardless of its classification accuracy. Explainability is therefore not merely a governance desideratum but a design constraint imposed by the legal environment of deployment.

A structural connection is also noted between the DoR standard and international data protection law. The GDPR's framework for automated decision-making—specifically the right not to be subject to solely automated processing under Article 22, combined with the transparency requirements of Articles 13–15 and the explanation right in Recital 71 [21]—applies to processing that produces "legal or similarly significant effects." FOIA/ATI withholding decisions plausibly satisfy this standard where the records relate to the requestor's own personal data. The proposed DoR standard is compatible with, and in several respects more operationally specific than, the GDPR's explainability requirements.

# 6 DISCUSSION AND LIMITATIONS

## 6.1 Generalizability Beyond the D.C. Circuit Corpus

The D.C. Circuit corpus provides unusually rich ground-truth signal—judicially reviewed Vaughn index entries with explicit criterion-level findings. However, this documentary richness is not uniformly distributed. National security and law enforcement exemptions are substantially over-represented relative to their share of total federal FOIA practice; routine Exemption 6 privacy withholdings in domestic regulatory agencies are under-represented. The D.C. Circuit has also developed an unusually detailed body of FOIA precedent; circuits with thinner precedent may require jurisdiction-specific ontological calibration.

A specific consequence of this distribution requires explicit acknowledgment. The largest NIPA performance gains reported in Section 4.3 appear on Exemptions 5 and 6—the two exemptions requiring the most nuanced multi-factor balancing. However, these are also the two exemptions whose litigated D.C. Circuit distribution diverges most from routine agency practice: Exemption 5 is heavily represented because deliberative process disputes are frequently litigated, while Exemption 6 routine privacy withholdings in domestic regulatory agencies are heavily under-represented. The headline precision figure of 0.847 should therefore be read as a performance estimate for the *litigated corpus distribution*, not for the full population of federal FOIA withholdings. Whether NIPA's gains on Exemptions 5 and 6 transfer to the routine, non-litigated agency context— where the documents are less formally structured and the exemption claims less sharply argued—is an open empirical question that requires a broader corpus to resolve.

Three steps are proposed to address the U.S. corpus limitation. First, expansion to incorporate Vaughn-index-equivalent records from other circuits, drawing on the National Security Archive's broader litigation database and agency Inspector General reports. Second, supplementation with records from state open-records litigation, requiring jurisdiction-specific LNO extensions but substantially expanding generalizability. Third, construction of a synthetic augmentation layer using retrieval-augmented generation over a

current corpus of FOIA opinions to produce additional annotated CWU examples for under-represented exemption types.

*International adaptability.* The LNO's modular structure supports extension to non-U.S. access-to-information regimes. Two illustrative mappings demonstrate this adaptability.

*EU Artificial Intelligence Act (Regulation (EU) 2024/1689).* The Act designates AI systems used by public authorities in administrative decision-making as high-risk under Annex III, subjecting them to conformity assessment, transparency, and human oversight obligations [6]. ATI withholding systems deployed by EU member state agencies would fall within this category. The LNO's criterion-level structure and the DoR standard are directly aligned with the Act's requirements for human oversight and logging of AI-assisted decisions: the criterion-level assessment record satisfies the Act's traceability obligations, and the oversight record satisfies its requirement that high-risk AI systems "allow for human oversight." A DoR-compliant NIPA deployment would therefore also satisfy the AI Act's transparency requirements for this use case, providing a unified compliance architecture for both ATI and AI governance obligations in EU contexts.

*Right to Information frameworks in the Global South.* Bangladesh's Right to Information Act 2009 is among the most progressive access-to-information statutes in South Asia, establishing a strong presumption of disclosure subject to a defined category of exemptions [5]. The exemption structure, while differing from the U.S. nine-exemption scheme in its formulation, is similarly norm-grounded and requires contextual balancing in several categories. The LNO's five-tuple formalization $\langle E, C, R, B, H \rangle$ is jurisdiction-agnostic at the structural level: adapting it to the Bangladesh RTI Act requires only re-specifying the exemption set $E$ and criterion mappings $R$, leaving the deontic modality framework and balancing relation $B$ intact. This suggests the framework is extensible to right-to-information constitutional and statutory regimes across South and Southeast Asia without requiring architectural changes to NIPA.

## 6.2 NIPA as Augmentation, Not Replacement

A natural evaluative question raised by reviewers concerns the relationship between NIPA and trained human FOIA reviewers. The inter-annotator agreement of $\kappa$=0.76 at the exemption level and $\kappa$=0.68 at the criterion level, achieved by the three-attorney annotation panel following the LNO-grounded protocol, reflects the level of agreement achievable by comparably instructed legal professionals. That NIPA's precision of 0.847 approaches—but does not reach—this upper bound is expected: no text-only model can replicate the contextual background knowledge, adversarial anticipation, or institutional memory of an experienced FOIA attorney.

The more important point, however, is that NIPA is not designed to replace human reviewers. It is designed to *structure and augment* their work in a manner that is legally compliant with the DoR standard. Three dimensions of this augmentation relationship are architecturally guaranteed by the LNO-NIPA framework. First, NIPA's criterion-level outputs function as a structured checklist for the human reviewer: rather than reading a document cold, the reviewer receives a pre-populated criterion assessment for each CWU, which they can accept, modify, or reject. This mirrors the role of structured checklists in high-stakes professional domains and reflects the general design principle that decision-support tools reduce reviewer error when they decompose complex multi-factor tasks into explicit sequential assessments. Second, the criterion-level rationale generated by NIPA directly produces the criterion-level assessment record required by the DoR standard's first element, reducing the documentation burden on the human reviewer without substituting for their judgment. Third, the oversight record required by the DoR standard's third element—documenting whether the AI recommendation was accepted, modified, or rejected—creates an institutional audit trail that makes human accountability explicit and enforceable.

A full human-reviewer comparison study—in which trained agency FOIA reviewers and NIPA independently process a common document set with criterion-level rationales collected from both—remains an important direction for future work, and would be particularly valuable for calibrating which CWU types benefit most from AI-assisted structuring versus those that require immediate escalation to unassisted human judgment.

## 6.3 Temporal Dynamics of Exemption Doctrine

The LNO as currently specified does not capture temporal dynamics in exemption doctrine. Judicial interpretations shift over time; the LNO would require systematic updates to remain current with evolving precedent. An automated ontology maintenance pipeline—perhaps using RAG over a current corpus of FOIA litigation—would address this limitation and constitutes an important direction for future work.

## 6.4 Structural Limitations of Text-Only Models

Exemption 1 withholdings often depend on information whose sensitivity derives entirely from context that is itself classified; the document may contain no internal signal of its sensitivity. This is a structural limitation that no prompt-engineering approach can overcome, and it reinforces the importance of the human-review element of the DoR standard.

## 6.5 Jurisdictional Scope of the DoR Standard

The DoR standard is grounded primarily in U.S. administrative law. Extension to Canada's ATI framework, the UK FOIA regime, and EU member state implementations of the Public Sector Information Directive would require jurisdiction-specific doctrinal analysis. The structural argument is likely to generalize, but the specific legal grounding requires adaptation.

# 7 CONCLUSION

This paper has argued that AI-assisted FOIA sensitivity review requires a fundamental reorientation: from classification to norm operationalization. The FOIA exemption scheme encodes computable legal norms—structured, multi-factor standards whose application requires reasoning over specific criteria, modalities, and balancing relationships. The Legal Norm Ontology and Norm-Informed Prompting Architecture developed here provide a framework for operationalizing these norms within LLM-based review pipelines,

demonstrating substantial performance improvements over unstructured baseline approaches. The Disclosure-of-Reasoning standard derived from administrative due process doctrine establishes that explainability in this domain is not merely a best practice but a legal requirement, and that systems architecturally incapable of generating criterion-level rationales cannot be lawfully deployed for government disclosure decisions.

The broader implications extend beyond FOIA. Wherever AI systems are deployed in contexts governed by legal standards that encode normative structure—benefits adjudication, immigration review, tax compliance—the argument for norm-informed architectures applies. Legal AI systems that ignore the normative structure of the standards they purport to apply are not merely less accurate; they are operating outside the legal framework they are supposed to instantiate. Bridging the gap between legal doctrine and AI system design is not only a technical challenge, but a condition of lawful governance.

## REFERENCES


[1] 1947. SEC v. Chenery Corp. 332 U.S. 194. Establishing that agency action must be sustained, if at all, on the grounds invoked by the agency at the time of decision.

[2] 1966. Freedom of Information Act. 5 U.S.C. § 552. As amended by the FOIA Improvement Act of 2016.

[3] 1983. Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co. 463 U.S. 29.

[4] 1989. DOJ v. Reporters Committee for Freedom of the Press. 489 U.S. 749.

[5] 2009. Right to Information Act 2009. Act No. XX of 2009, Government of the People's Republic of Bangladesh.

[6] 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). O.J. L, 12 July 2024. Annex III (high-risk AI systems in public administration).

[7] Bayan Altalla', Sameera Abdalla, Ahmad Altamimi, Layla Bitar, Amal Al Omari, Ramiz Kardan, and Iyad Sultan. 2025. Evaluating GPT Models for Clinical Note De-Identification. *Scientific Reports* 15, 1 (2025), 3852.

[8] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. arXiv preprint arXiv:2401.13018.

[9] Guido Governatori and Antonino Rotolo. 2004. On the Axiomatics of New Normative Systems. *Logique et Analyse* 186 (2004), 153–178.

[10] Maura R. Grossman and Gordon V. Cormack. 2011. Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law and Technology* 17, 3 (2011), 1–48.

[11] Louis Kaplow. 1992. Rules versus Standards: An Economic Analysis. *Duke Law Journal* 42, 3 (1992), 557–629.

[12] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Shen, Dinggang Li, and Tianming Chen. 2023. DeID-GPT: Zero-Shot Medical Text De-Identification by GPT-4. arXiv preprint arXiv:2303.11032.

[13] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2020. Active Learning Stopping Strategies for Technology-Assisted Sensitivity Review. In *Proceedings of the 43rd International ACM SIGIR Conference (SIGIR '20)*. ACM, New York, NY, USA, 1379–1388.

[14] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2020. How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review. *ACM Transactions on Information Systems* 39, 1 (2020), 1–34.

[15] Monica Palmirani, Guido Governatori, and Corrado Roversi. 2021. Legal Ontologies for the Semantic Web. In *Intelligent Systems for Law*. Springer.

[16] Livio Robaldo, Jacob Leibowitz, Johan De Smedt, and Luigi Di Caro. 2020. Introduction to the Special Issue on AI and Law. *Artificial Intelligence and Law* 28 (2020), 1–12.

[17] Marek J. Sergot, Fariba Sadri, Robert A. Kowalski, Frank Kriwaczek, Peter Hammond, and H. T. Cory. 1986. The British Nationality Act as a Logic Program. *Commun. ACM* 29, 5 (1986), 370–386.

[18] Richard E. Susskind. 1987. *Expert Systems in Law: A Jurisprudential Inquiry*. Oxford University Press.

[19] U.S. Department of Justice. 2024. *Annual Report to the Attorney General on Freedom of Information Act Administration for Fiscal Year 2023*. Technical Report. Office of Information Policy.

[20] Georg Henrik von Wright. 1951. Deontic Logic. *Mind* 60, 237 (1951), 1–15.

[21] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–887.

[22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 24824–24837.