ComplianceNLP: Knowledge-Graph-Augmented RAG for Multi-Framework Regulatory Gap Detection

Dongxin Guo[1], **Jikun Wu**[2], **Siu Ming Yiu**[1]
[1]The University of Hong Kong [2]Stellaris AI Limited
bettyguo@connect.hku.hk, hk950014@connect.hku.hk, smyiu@cs.hku.hk

Abstract

Financial institutions must track over 60,000 regulatory events annually across fragmented jurisdictions, a volume that overwhelms manual compliance teams. The industry has paid over $300 billion in fines and settlements since the 2008 financial crisis. We present COMPLIANCENLP, an end-to-end system that automatically monitors regulatory changes, extracts structured obligations, and identifies compliance gaps against institutional policies. The system integrates three components: (1) a knowledge-graph-augmented RAG pipeline grounding generations in a regulatory knowledge graph of 12,847 provisions across SEC, MiFID II, and Basel III; (2) multi-task obligation extraction combining NER, deontic classification, and cross-reference resolution over a shared LEGAL-BERT encoder; and (3) compliance gap analysis that maps obligations to internal policies with severity-aware scoring. On our benchmark, COMPLIANCENLP achieves 87.7 F1 on gap detection, outperforming GPT-4o+RAG by +3.5 F1, with 94.2% grounding accuracy (r=0.83 vs. human judgments) and 83.4 F1 under realistic end-to-end error propagation. Ablations show that knowledge-graph re-ranking contributes the largest marginal gain (+4.6 F1), confirming that structural regulatory knowledge is critical for cross-reference-heavy tasks. Domain-specific knowledge distillation (70B→8B) combined with Medusa speculative decoding yields 2.8× inference speedup, enabled by an empirical finding that regulatory text's low entropy (H=2.31 bits vs. 3.87 general text) produces 91.3% draft-token acceptance rates. In four months of parallel-run deployment processing 9,847 updates at a financial institution, the system achieved 96.0% estimated recall and 90.7% precision, with a 3.1× sustained analyst efficiency gain. We report deployment lessons spanning trust calibration, GRC integration, and distributional shift monitoring that inform future regulated-domain NLP deployments.

1 Introduction

Compliance with financial regulation has become a substantial operational task for banks. A single multinational bank may be subject to hundreds of regulatory frameworks across dozens of jurisdictions (Du et al., 2025). Over 60,000 regulatory events were tracked across 190 countries in 2022 (Thomson Reuters Regulatory Intelligence, 2023), far exceeding the capacity of human compliance teams. Global banks have paid over $300 billion in regulatory fines and settlements since the 2008 financial crisis (Boston Consulting Group, 2017).

Despite clear demand, automated regulatory compliance remains underserved within NLP. Legal NLP has advanced benchmarks and domain-adapted models (Guha et al., 2023; Chalkidis et al., 2022, 2020), and the COLING 2025 Regulations Challenge (Gokhan et al., 2025) highlighted significant gaps in LLM performance on financial regulatory tasks. Yet no end-to-end system bridges these advances to deliver production-ready compliance monitoring; commercial GRC platforms such as Ascent RegTech and Wolters Kluwer OneSumX still rely heavily on rule-based approaches with manual curation (Jain et al., 2025) (Appendix A).

We introduce COMPLIANCENLP, a system that automatically monitors regulatory changes, extracts structured obligations, and identifies compliance gaps against internal policies. The system currently covers three major frameworks (SEC, MiFID II, Basel III), representing approximately half of the annual regulatory update volume. Our approach addresses three challenges: (1) *regulatory grounding*: LLMs are prone to hallucination (Wang et al., 2023), which we mitigate through KG-augmented RAG achieving 94.2% grounding accuracy (r=0.83 vs. human; §6.1); (2) *multi-framework obligation extraction*: prior systems target single frameworks (Cejas et al., 2023; Bayer et al., 2025), while we jointly handle three; and

(3) *production-grade latency*: we achieve sub-second p50 inference via knowledge distillation (Gu et al., 2024) and Medusa speculative decoding (Cai et al., 2024).

The system achieves 87.7 gap detection F1 at the evaluation threshold (δ=0.6) and 85.3 macro F1 at the recall-optimized deployment threshold (δ=0.45), where Full Gap recall reaches 91.2%. Our contributions are:

1. An end-to-end regulatory compliance monitoring system integrating KG-augmented RAG, multi-task obligation extraction, and automated gap analysis across three frameworks, with 83.4 F1 under realistic error propagation and two independent additive analyses isolating each module's contribution.
2. A Regulatory Knowledge Graph linking 12,847 provisions with 94.7% edge precision and 87.3% estimated recall.
3. A deployment analysis including production metrics from 4 months of parallel-run operation (9,847 updates, estimated 96.0% recall), cost modeling, and five lessons learned.
4. A production optimization pipeline where domain-specific distillation (70B→8B) and Medusa speculative decoding yield 2.8× combined speedup, with the finding that regulatory text's low entropy enables 91.3% Medusa acceptance rates.

2 Related Work

Legal and regulatory NLP. Legal NLP benchmarks including LegalBench (Guha et al., 2023), LexGLUE (Chalkidis et al., 2022), and CUAD (Hendrycks et al., 2021), together with domain models such as LEGAL-BERT (Chalkidis et al., 2020) and the MultiLegalPile (Niklaus et al., 2024), provide foundations for legal language understanding (Ariai et al., 2025; Leitner et al., 2020). For regulatory compliance specifically, Doc2Doc retrieval (Chalkidis et al., 2021) identified limitations of text similarity for legislation, ObliQA/RIRAG (Bayer et al., 2025) introduced regulatory QA, and DERECHA (Cejas et al., 2023) demonstrated GDPR compliance checking at 89.1% precision but assumes pre-structured policy clauses and targets a single framework, whereas ComplianceNLP jointly handles three frameworks end-to-end from raw regulatory text and reports comparable precision (90.7%) at production scale (Table 4). Most closely related, Sun et al. (2025) propose a RAG-based compliance checker built on an eventic graph; we differ in (i) spanning three regulatory frameworks rather than a single corpus, (ii) grounding retrieval in a typed regulatory knowledge graph rather than purely embedding-based vector store, (iii) learning end-to-end obligation extraction rather than assuming structured regulatory inputs, and (iv) reporting four months of parallel-run deployment evidence. Sancheti et al. (2022) addressed deontic modality classification. None handles multi-framework obligation extraction with production deployment.

Financial LMs, RAG, and efficient inference. Financial domain models (Araci, 2019; Wu et al., 2023; Yang et al., 2023; Xie et al., 2023; Nie et al., 2024) provide domain adaptation foundations. RAG (Lewis et al., 2020) grounds LLM outputs, with KG-enhanced variants improving factual QA (Baek et al., 2023), multi-hop reasoning (Sen et al., 2023), and corpus-level queries (Edge et al., 2024). LegalBench-RAG (Pipitone and Alami, 2024) and CLERC (Hou et al., 2025) benchmark legal retrieval, and LexKeyPlan (T.y.s.s and Hernandez, 2025) augments legal-text generation with anticipatory keyphrase planning to align retrieval with intended future content. For hallucination mitigation, MiniCheck (Tang et al., 2024) offers efficient fact-checking, the FACTS benchmark (Jacovi et al., 2025) evaluates long-document factuality, CoCoLex (T.y.s.s et al., 2025) enforces faithfulness via confidence-guided copy-based decoding for grounded legal generation, and Kenthapadi et al. (2024) report industry grounding lessons. For efficient inference, Medusa (Cai et al., 2024) extends speculative decoding (Leviathan et al., 2023; Chen et al., 2023) with multiple heads, and MiniLLM (Gu et al., 2024) enables effective distillation. Most relevant, Zagyva et al. (2025) demonstrated Medusa+KD at Booking.com, a setup we adapt for compliance with domain-specific findings on acceptance rates.

3 The ComplianceNLP System

Figure 1 presents the pipeline: (i) ingestion and indexing of regulatory documents into the RKG and vector store; (ii) multi-task obligation extraction; and (iii) compliance gap analysis mapping extracted obligations against internal policies. The system uses two distinct model families sequentially: a LEGAL-BERT-based extraction module

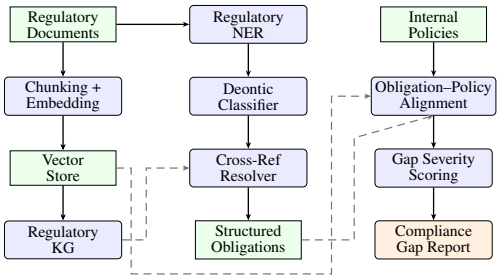

Figure 1: Overview of the COMPLIANCENLP pipeline: regulatory documents are ingested into a vector store and knowledge graph (left), processed by multi-task obligation extraction (center), and compared against internal policies to produce gap reports (right). Dashed arrows indicate cross-module data flow.

(§3.2) produces structured obligations, which feed the LLaMA-3-based generator (§3.3) for gap analysis. The two models share no parameters. Full pseudocode and notation are in Appendix B.

3.1 Regulatory RAG Pipeline

Hybrid retrieval. Given a regulatory query q, we retrieve candidate passages using a weighted combination of dense and sparse scores: $s(q,d) = \alpha \ \text{sim}_{\text{dense}}(q,d) + (1-\alpha) \ \text{BM25}(q,d)$, where α=0.7 and the dense encoder is a legal-domain bi-encoder fine-tuned from `all-MiniLM-L6-v2` on 50K regulatory passage pairs. Top-k passages (k=5) are re-ranked using KG proximity: $s_{\text{KG}}(q,d) = \beta \ \text{KGScore}(q,d,\mathcal{G}) + (1-\beta) \ s(q,d)$, where β=0.3 and KGScore measures the graph distance between the query's source provision and retrieved passages' linked provisions.

RKG construction. The RKG is constructed via a semi-automated pipeline using three format-specific parsers for SEC EDGAR XML, EUR-Lex HTML, and BIS PDF. Entity and obligation nodes are extracted using fine-tuned NER models (§3.2). Cross-reference edges are identified using regex patterns combined with a learned span-pair linker (a bilinear classifier over source and target provision embeddings), achieving 91.8% accuracy on 500 held-out pairs. A compliance expert reviewed 500 randomly sampled edges, yielding precision of 94.7%. Two experts enumerated all cross-references from 150 randomly sampled provisions, yielding estimated recall of 87.3% $\pm$ 3.8% (95% CI). The KG contains 12,847 provision nodes and 34,219 edges, updated nightly from regulatory feeds with 18-hour median incorporation latency. Full schema details are in Appendix D.

Update lifecycle and consistency. Real-world regulatory monitoring requires the RKG to remain current as new provisions are published, while preserving consistency during partial updates. The system synchronises nightly with SEC EDGAR and EUR-Lex RSS feeds; emergency or out-of-cycle updates can be triggered manually or flagged automatically when confidence-score anomalies appear on incoming documents. During the parallel run, a Basel III amendment introducing nested conditional obligations was detected within 6 hours via this anomaly signal and resolved through targeted retraining within 48 hours (Appendix O). On ingestion, new provisions are inserted into the RKG incrementally as node and edge writes; affected subgraph embeddings are recomputed within a 2-hop neighbourhood of the insertion site rather than over the full graph, and the FAISS vector store is rebuilt nightly with approximate nearest-neighbour insertion used between rebuilds. During the maximum 18-hour blind spot before the next synchronisation, the system falls back to embedding-only retrieval, whose ablation cost we report as -4.6 gap F1 (Table 2, "w/o KG reranking"). Provisions added since the last synchronisation carry a consistency flag that excludes them from KG re-ranking until validation completes. For time-critical announcements such as emergency SEC orders, this embedding-only fallback is the expected operating mode rather than a degraded one.

3.2 Multi-task Obligation Extraction

Regulatory documents express obligations through deontic modalities (*shall*, *must*, *may not*), nested cross-references, and domain-specific entities (Sancheti et al., 2022). We formulate obligation extraction as a multi-task problem with three jointly trained heads over a shared LEGAL-BERT encoder (Chalkidis et al., 2020) further fine-tuned on Pile of Law (Henderson et al., 2022): (a) a CRF layer for regulatory NER across 23 entity types[1], extending Leitner et al. (2020) with finance-specific types from FiNER (Loukas et al., 2022);

[1]Our schema goes beyond conventional person/organisation/location NER. It defines regulatory-domain role types (e.g., REGULATED_ENTITY for "investment firm", REPORTING_ENTITY for "registrant", SUPERVISORY_AUTHORITY), together with finance-specific span types (e.g., FINANCIAL_INSTRUMENT, THRESHOLD_VALUE, CAPITAL_REQUIREMENT, COMPLIANCE_PERIOD, JURISDICTION). These roles are defined by the regulatory frameworks themselves rather than by traditional named-entity ontologies. Full inventory in Appendix J.

(b) sentence-level deontic modality classification into OBLIGATION, PERMISSION, PROHIBITION, and RECOMMENDATION; and (c) a span-pair classifier for cross-reference resolution, linking to target provisions in the RKG. The combined loss is $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{NER}} + \lambda_2 \mathcal{L}_{\text{deontic}} + \lambda_3 \mathcal{L}_{\text{xref}}$ with λ_1=0.4, λ_2=0.3, λ_3=0.3.

Training uses 8,742 annotated regulatory sentences spanning SEC (3,211), MiFID II (2,987), and Basel III (2,544); the held-out evaluation set of 1,847 sentences (REGOBLIGATION, §5) was annotated by the same three compliance experts using identical guidelines, with leakage prevented by document-level partitioning rather than sentence-level random splits. The aggregate inter-annotator agreement is κ=0.84 (Fleiss); per-layer agreement is highest for deontic classification (κ=0.89) and NER (κ=0.86), and lowest for cross-reference resolution (κ=0.78), where ambiguous implicit references that lack explicit citation markers are the dominant disagreement source. Training data is augmented with silver-standard annotations from ObliQA (Bayer et al., 2025) and COLING 2025 shared task data (Gokhan et al., 2025) (zero overlap with the manually-annotated evaluation splits). Annotation guidelines and the full entity-type inventory are in Appendix J.

3.3 Compliance Gap Analysis

Each extracted obligation o_j is represented as $\langle$*entity*, *action*, *modality*, *condition*, *source_provision*$\rangle$. We embed both obligations and internal policy clauses and compute alignment scores: $a(o_j, p_k) = \text{sim}_{\text{dense}}(o_j, p_k) \cdot f_{\text{type}}(o_j, p_k)$, where f_{type} is a learned fuzzy type-matching function handling naming convention differences (e.g., "credit institution" vs. "bank"). Policy clauses with maximum alignment below δ=0.6 (evaluation) or δ=0.45 (recall-optimized deployment) are flagged as potential gaps and classified into COMPLIANT, PARTIAL GAP, or FULL GAP by the generator LLM, conditioned on obligation modality, entity types, and enforcement history from the RKG. The system produces structured gap reports with source citations, severity classification, and remediation recommendations. We analyze threshold sensitivity in §6.

4 Production Optimization

Real-time compliance monitoring demands low-latency inference. End-to-end document processing (4.2 min/doc) includes extraction, retrieval, and gap analysis; "real-time" refers here to the generator LLM's inference latency constraint.

Config.	p50 (ms)	p99 (ms)	Speed up	Acc. Retention (%) NER	Gap	QA	Grd.
70B Teacher	1847	3214	1.0×	100	100	100	100
8B (SFT only)	897	1492	2.1×	95.1	95.4	94.8	95.2
8B (KD only)	824	1387	2.2×	96.8	97.0	96.3	96.8
8B+Med. (gen.)	793	1295	2.3×	96.4	96.7	96.1	96.5
8B+Med. (dom.)	**659**	**1082**	**2.8×**	**98.6**	**98.1**	**97.4**	**98.3**

Table 1: Latency and accuracy trade-offs for the distilled student model *in isolation* (without system components). KD achieves numerically higher retention than SFT (96.8% vs. 95.1% NER), though CIs overlap. Domain Medusa heads retain ≥97.4% across all tasks. The full system's 91.3 NER F1 (Table 2) reflects additional gains from KG re-ranking, multi-task training, and MiniCheck.

We distill compliance capabilities from LLaMA-3-70B-Instruct (fine-tuned on regulatory data) into LLaMA-3-8B-Instruct following MiniLLM's reverse KL divergence approach (Gu et al., 2024): $\mathcal{L}_{\text{KD}} = \text{KL}(p_{\text{student}} \| p_{\text{teacher}}) + \gamma\, \mathcal{L}_{\text{SFT}}$ with γ=0.5, trained on 15K compliance instruction–response pairs. KD alone provides 2.2× speedup (1,847→824 ms p50). We augment the student with M=3 Medusa prediction heads (Cai et al., 2024) trained on 2.1M regulatory tokens, extending the Medusa+KD paradigm of Zagyva et al. (2025). Regulatory language's constrained vocabulary yields lower entropy (H=2.31 bits vs. 3.87 on C4 (Raffel et al., 2020)), enabling 91.3% token acceptance vs. 82.7% for general-text heads. Combined speedup: 2.8× (659 ms p50). Table 1 reports the full accuracy–latency breakdown.

5 Experiments

Datasets. REGOBLIGATION: 1,847 regulatory sentences from SEC (712), MiFID II (614), and Basel III (521), manually annotated for 23 NER entity types, four deontic modalities (OBLIGATION, PERMISSION, PROHIBITION, RECOMMENDATION), and cross-references (to be released). GAPBENCH: 423 obligation–policy pairs from a single financial institution (COMPLIANT: 210, PARTIAL GAP: 128, FULL GAP: 85; κ=0.81; 95% CI for Full Gap F1: ±4.2). We also evaluate on ObliQA (5,574 questions) (Bayer et al., 2025) and the COLING 2025 Challenge (312 questions) (Gokhan et al., 2025). Dataset details and release plans are in Appendix C.

System	Obligation		Gap	Reg. QA	
	NER	Deon.	Det.	EM	F1
Without retrieval					
LEGAL-BERT	82.1	84.6	71.3	–	–
FinBERT	76.4	79.2	68.1	–	–
GPT-4 (5-shot)	84.7	87.3	79.8	41.2	58.6
GPT-4o (5-shot)	85.9	88.1	81.4	43.7	61.3
With retrieval					
RIRAG	–	–	–	38.9	54.2
GPT-4 + RAG	87.4	89.8	82.9	46.3	64.7
GPT-4o + RAG	88.6	90.5	84.2	48.1	66.8
LLaMA-3-8B + RAG	87.9	89.8	83.5	47.4	65.9
LLaMA-3-70B*	90.2	91.8	86.3	49.1	67.4
COMPLIANCENLP	**91.3**†‡	**92.7**†‡	**87.7**†‡	**52.8**†‡	**71.9**†‡
w/o KG reranking	88.4	91.2	83.1	47.6	66.2
w/o multi-task	89.1	90.4	84.9	50.3	69.1
w/o MiniCheck	91.0	92.5	87.2	52.1	71.0

Table 2: Main results (F1) at evaluation threshold δ=0.6. *Teacher (pre-distillation, full RAG). †Significant vs. LLaMA-3-70B (p<0.05, paired bootstrap). ‡Significant vs. GPT-4o+RAG (p<0.05). GPT-4/4o+RAG use our hybrid retrieval without KG/multi-task/MiniCheck. Std in Appendix Table 10. See Table 3 for deployment threshold.

	Eval (δ=0.6)			Deploy (δ=0.45)		
Class	P	R	F1	P	R	F1
Compliant	92.1	94.3	93.2	88.4	96.8	92.4
Partial Gap	85.4	82.1	83.7	78.1	87.3	82.4
Full Gap	83.2	79.8	81.5	72.8	91.2	81.0
Macro avg.	86.9	85.4	86.1	79.8	91.8	85.3

Table 3: Per-class gap detection at both thresholds. The deployment threshold (δ=0.45) prioritizes Full Gap recall (91.2%) at the cost of lower precision, yielding similar Full Gap F1 at both thresholds. 95% CI for Full Gap F1: ±4.2 (bootstrap). Manual analysis of 17 missed Full Gaps: implicit obligations (35%), multi-hop cross-references (29%), jurisdiction-specific nuance (21%). See Appendix F for details.

Baselines. *Without retrieval*: GPT-4 and GPT-4o (5-shot), LEGAL-BERT (Chalkidis et al., 2020), FinBERT (Araci, 2019). *With retrieval*: GPT-4+RAG and GPT-4o+RAG (same hybrid retrieval, no KG/multi-task/MiniCheck), LLaMA-3-8B+RAG (distilled student without domain components), LLaMA-3-70B-Instruct (teacher, full RAG), and RIRAG (Bayer et al., 2025). All results report mean over 3 seeds; significance via paired bootstrap (n=10,000) (Koehn, 2004).

6 Results and Analysis

Main results. Table 2 shows results at δ=0.6. COMPLIANCENLP achieves 91.3 NER F1, 87.7 gap detection F1, and 94.2% grounding accuracy, compared to 83.7% for GPT-4+RAG and 85.1% for GPT-4o+RAG (both without post-generation verification). Compared to GPT-4o+RAG (same retrieved passages), we improve by +2.7 NER F1, +2.2 Deontic F1, +3.5 gap detection F1, and +5.1 QA F1, all statistically significant (p<0.05). Compared to LLaMA-3-8B+RAG (same student model without domain components), improvements of +3.4 NER F1 and +4.2 gap F1 confirm that performance gains are attributable to domain-specific components rather than model family effects. Domain-specific baselines in Table 2 also show large absolute margins: LEGAL-BERT (Chalkidis et al., 2020), the standard legal NLP encoder, reaches 82.1 NER F1 and 84.6 deontic F1 vs. our 91.3 / 92.7 (+9.2 / +8.1), and RIRAG (Bayer et al., 2025), a regulatory-domain QA system, reaches 54.2 QA F1 vs. our 71.9 (+17.7), with statistical significance per the bootstrap test.

Ablation. Removing KG re-ranking causes the largest drop (−4.6 gap F1), confirming structural regulatory knowledge is essential. Removing multi-task training reduces NER by 2.2 F1. MiniCheck removal has minimal F1 impact but degrades grounding accuracy from 94.2% to 86.7%, since it filters unfaithful justifications without changing gap classification.

Additive component analysis. Starting from GPT-4o+RAG, incrementally adding KG re-ranking (+2.2 gap F1), multi-task extraction (+0.7), and MiniCheck (+0.6) reconstructs the full system. A complementary analysis from LLaMA-3-8B+RAG yields consistent rankings (KG: +2.5; multi-task: +0.8; MiniCheck: +0.9), confirming attributions are robust to base model choice (Appendix Tables 8, 9).

Grounding and end-to-end evaluation. MiniCheck grounding accuracy (94.2%) is validated against human annotations (r=0.83, κ=0.87; §6.1), with accuracy declining for complex cross-references (97.1% at 0 refs → 84.6% at 3+). Running the full pipeline on 150 held-out documents, extraction errors propagate in 12.3% of cases, reducing F1 from 87.7 to 83.4, with NER boundary errors as the primary source (68%). The 83.4 end-to-end F1 translates to approximately 2.1 missed obligations per 100-page regulatory document and 1.3 false gap alerts per day at

current processing volume. At the deployment threshold (δ=0.45), false alerts increase to 3.7/day while missed obligations drop to 0.7/document, a trade-off analysts find acceptable (§7).

Error analysis. Manual examination of 100 obligation extraction errors reveals three dominant categories: boundary errors in long nested entities (34%), cross-reference resolution failures for multi-document references (28%), and ambiguous deontic modality where regulatory language uses hedged phrasing (19%). Performance varies markedly across the three frameworks: SEC texts (n=712) achieve the highest NER F1 (93.1) and cross-reference F1 (89.7) due to standardised filing formats parsed directly from EDGAR XML; MiFID II (n=614) sits in the middle (NER 91.4, XRef 85.3, reflecting the mix of directive text and delegated-regulation cross-references); Basel III (n=521) is the most challenging, with cross-reference F1 dropping to 81.2 because nested multi-level references span accord revisions (e.g., d295 → d424 → CRR Art. 412). Grounding accuracy follows the same ordering (95.8 / 94.1 / 92.4 across SEC / MiFID II / Basel III), and Basel III documents accordingly receive mandatory analyst review in deployment (Appendix H, Table 11).

6.1 Grounding Metric Validation

Two compliance experts independently annotated 200 randomly sampled outputs for grounding accuracy (IAA κ=0.87). MiniCheck achieved 91.8% agreement with human majority labels (precision: 93.2%, recall: 90.4%; Pearson r=0.83). Of 16 system–vs–human disagreements, 62% involved provision confusion (e.g., conflating Art. 25(1) with Art. 25(2)), 25% subtle modality errors (*should* vs. *shall*), and 13% incomplete citation. Grounding accuracy degrades smoothly with cross-reference complexity: 97.1% on obligations with 0 references, 93.4% at 1, 89.8% at 2, and 84.6% at 3+; on nested conditional obligations spanning multiple articles, accuracy drops further to an estimated ∼79% (n=38). This degradation pattern motivates the deployment policy of mandatory analyst review for High and Critical findings whose source obligation involves ≥3 cross-references. Full results are in Appendix E.

7 Deployment and Lessons Learned

COMPLIANCENLP is in parallel-run deployment at a financial institution, operating alongside existing manual compliance processes.[2] Deployment details including timeline, integration architecture, and monthly trend analysis are in Appendix K.

Metric	Value
Regulatory updates processed	9,847
Potential gaps flagged	1,243
Gaps confirmed by analyst review	1,127 (90.7%)
Gaps missed, caught by manual	47
Estimated production recall	**96.0%**
Production precision	**90.7%**
Analyst override rate	8.3%
System uptime	99.2%
Median end-to-end latency	4.2 min/doc

Table 4: Operational metrics from 4 months of parallel-run deployment (Oct 2025 – Jan 2026, δ=0.45). Recall = gaps caught / (caught + missed); see Appendix L for estimation caveats and sensitivity analysis.

Key deployment findings. Over 4 months of parallel operation on 9,847 regulatory updates: (1) estimated 96.0% production recall at δ=0.45 with 90.7% precision (Table 4); (2) 3.1× sustained analyst efficiency gain (47→15 min/update, p<0.01, paired t-test, n=96); (3) monotonically improving monthly performance, with precision rising 88.2% → 92.4% and recall 94.1% → 97.3% as the analyst override rate fell from 11.4% to 6.1% over the 4 months; (4) the system identified 24 confirmed gaps the manual process had not yet flagged. Current Phase 2 cost is $48.2K/month against a manual baseline of $45K (+7%, since all outputs still receive human review); the projected Phase 3 and Phase 4 costs of $33.2K and $20.2K (savings of 26% and 55%) emerge as graduated autonomy reduces the review burden on low-severity findings (Appendix M).

User study. During Phase 2, 12 compliance analysts processed 96 regulatory updates (8 per analyst) both manually and with system assistance (counterbalanced order, unblinded). System-assisted review reduced average processing time from 47 to 12 minutes (3.9×, p<0.01, paired t-test). After the controlled study, processing time stabilized at 15 min/update (3.1×), confirming sustained efficiency gains after novelty attenuation.

Cross-institutional generalization. A key concern is whether COMPLIANCENLP generalizes beyond our partner institution. On 87 obligation–policy pairs from a European partner bank, the

[2] All system outputs are reviewed by analysts; disagreements are logged for continuous improvement.

system achieved 84.1 gap detection F1 *without* institution-specific fine-tuning (a 3.6-point drop from our primary evaluation). After fine-tuning f_{type} on just 50 examples, performance recovered to 86.3 F1, with adaptation gains concentrated on the harder gap classes (Compliant 89.2→91.4, Partial Gap 79.8→82.1, Full Gap 78.4→82.7 F1), the classes that drive operational value. A controlled perturbation experiment replacing institutional policies with synthetically perturbed variants supports our estimate that ∼70% of difficulty is regulation-specific and ∼30% institution-specific. Full cross-institutional evaluation (∼1,200 examples) is expected by Q3 2026. Details in Appendix N.

7.1 Key Lessons

1. Structural knowledge outperforms embeddings for cross-references. Embedding-only prototypes achieved 72.3 F1 on cross-reference resolution; adding KG-based structural relationships improved this to 89.1 F1 (+16.8). This was the single most impactful design decision.

2. Formulaic language enables efficient speculative decoding. Regulatory text's constrained vocabulary yields 91.3% Medusa acceptance rates vs. 82.7% on general text (H=2.31 vs. 3.87 bits), suggesting speculative decoding may be particularly effective in other low-entropy domains (medical records, patent filings).

3. Analysts trust recall more than F1. A single missed compliance gap eroded confidence significantly. Unlike clinical NLP where false negatives carry individual risk, compliance false negatives carry *institutional* risk, making analysts even more loss-averse. This motivated recall-optimized thresholds and prominent confidence score display.

4. GRC integration is harder than model development. Integrating COMPLIANCENLP into the institution's existing GRC platform, which used a 15-year-old, 847-category compliance taxonomy organised by regulatory source while expecting outputs categorised by business function, consumed roughly three months of engineering, comparable to the entire model development cycle. We recommend that future deployments budget at least 40% of total project timeline for GRC integration and begin taxonomy mapping in parallel with model development rather than sequentially (Appendix O).

5. Organisational adoption requires staged trust-building. In Month 1, system-assisted processing time was 22 min/update because analysts re-verified 78% of system outputs; by Month 4 this dropped to 12 min/update with a 23% re-verification rate, concentrated on high-severity findings. The most effective intervention was publishing weekly "system vs. manual" accuracy digests that transparently reported both successes and failures, allowing analysts to track cumulative evidence rather than rely on point demonstrations.

8 Conclusion

We presented COMPLIANCENLP, an end-to-end system for automated regulatory compliance monitoring combining KG-augmented RAG, multi-task obligation extraction, and compliance gap analysis, optimized for production via domain-specific knowledge distillation. Two independent additive analyses demonstrate that KG re-ranking provides the largest marginal gain on gap detection (+2.2/+2.5 F1), confirming structural regulatory knowledge is essential for cross-reference-heavy tasks. The system achieves 87.7 F1 (δ=0.6), 83.4 under realistic error propagation, and 94.2% grounding accuracy. At the recall-optimized deployment threshold (δ=0.45), the system achieves an estimated 96.0% production recall and 90.7% precision over 4 months of parallel operation processing 9,847 updates, with a 3.1× sustained analyst efficiency gain, and preliminary cross-institutional evaluation (84.1→86.3 F1 after minimal adaptation) providing initial evidence of transferability. Five deployment lessons cover KG re-ranking effectiveness, constrained-vocabulary speculative decoding, recall-driven trust dynamics, GRC integration challenges, and organizational trust-building. Together they offer actionable guidance for practitioners building regulated-domain NLP systems. Future work includes extending to additional frameworks (targeting full 60K+ annual coverage), completing the multi-institutional GapBench expansion, and transitioning through later deployment phases toward full production.

Reproducibility. All materials are available at: https://github.com/bettyguo/ComplianceNLP.

Limitations

Our work has several limitations. First, COMPLIANCENLP covers three regulatory frameworks (∼48% of annual updates); extending to full coverage requires additional parser engineering. Second,

our primary evaluation dataset GAPBENCH (423 examples) derives from a single institution, with wide per-class confidence intervals (Full Gap F1 = 81.5 $\pm$ 4.2). Third, the system operates on English-language texts only. Fourth, our user study (12 analysts, 96 updates) is modest and unblinded, which may affect quality assessments. Fifth, production metrics are from parallel-run operation where all outputs receive human review; performance under reduced oversight (Phases 3–4) remains unverified. Sixth, the 96.0% production recall is an estimate with irreducible structural uncertainty: both the system and manual process may share blind spots, and the correlation between system confidence and manual detection difficulty is weak (r=0.18, p=0.07, n=47); under sensitivity analysis, if the manual process has 90–95% recall, estimated system recall adjusts to 95.6–95.8%. Seventh, the p99 latency of 1,082 ms exceeds our sub-second target. Additional limitations are discussed in Appendix P.

Ethical Considerations

Automated compliance monitoring raises important ethical considerations. Our system is a decision-support tool that augments human professionals; all high-severity findings require human review with full audit trails. The system should not be used as the sole basis for compliance decisions.

Data privacy is paramount: all regulatory texts used for training and evaluation are publicly available, and internal policy documents were anonymized before any model training. The system processes only regulatory texts and institutional policies; it does not handle personal customer data. We identify automation bias as a key risk: during Phase 2 deployment, we observed that 34% of analysts initially accepted system outputs without independent verification, which decreased to 12% after implementing weekly calibration exercises and confidence score visualization. The system's structured output format and mandatory confidence thresholds are designed to encourage, not replace, human judgment.

We acknowledge that deploying such systems could lead to workforce reductions. Our partner institution has committed to redeploying affected staff to higher-value compliance analysis rather than reducing headcount. We urge organizations deploying similar systems to prioritize workforce transition planning.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, and The University of Hong Kong and Stellaris AI Limited for their support of this work.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint*, arXiv.1908.10063.

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).

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint*, arXiv.2306.04136.

Ozan Bayer, Elif Nehir Ulu, Yasemin Sarkin, Ekrem Sütçü, Defne Buse Çelik, Alper Karamanlioglu, Ismail Karakaya, and Berkan Demirel. 2025. A REG-NLP framework: Developing retrieval-augmented generation for regulatory document analysis. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025 - Workshops, Abu Dhabi, UAE, January 19-24, 2025*, pages 97–101. Association for Computational Linguistics.

Boston Consulting Group. 2017. Global risk 2017: Staying the course in banking. Technical report, Boston Consulting Group. https://www.bcg.com/publications/2017/financial-institutions-growth-global-risk-2017-staying-course-banking.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 5209–5235. PMLR / OpenReview.net.

Orlando Amaral Cejas, Muhammad Ilyas Azeem, Sallam Abualhaija, and Lionel C. Briand. 2023. Nlp-based automated compliance checking of data processing agreements against gdpr. *IEEE Transactions on Software Engineering*, 49(9):4282–4303.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *arXiv preprint*, arXiv.2010.02559.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021. Regulatory compliance through doc2doc information retrieval: A case study in EU/UK legislation where

text similarity has limitations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3498–3511. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4310–4330. Association for Computational Linguistics.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint*, arXiv.2302.01318.

Kelvin Du, Yazhi Zhao, Rui Mao, Frank Xing, and Erik Cambria. 2025. Natural language processing in finance: A survey. *Inf. Fusion*, 115:102755.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint*, arXiv.2404.16130.

Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2025. Proceedings of the 1st regulatory NLP workshop (regnlp 2025). In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025 - Workshops, Abu Dhabi, UAE, January 19-24, 2025*. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, K. Aditya, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: an expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn J. Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. CLERC: A dataset for u. s. legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 7898–7913. Association for Computational Linguistics.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, and 7 others. 2025. The FACTS grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint*, arXiv.2501.03200.

Jivitesh Jain, Nivedhitha Dhanasekaran, and Mona T. Diab. 2025. From complexity to clarity: Ai/nlp's role in regulatory compliance. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 26629–26641. Association for Computational Linguistics.

Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. 2024. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6523–6533. ACM.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395. ACL.

Elena Leitner, Georg Rehm, and Julián Moreno Schneider. 2020. A dataset of german legal documents for named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4478–4485. European Language Resources Association.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. Finer: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4419–4431. Association for Computational Linguistics.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint*, arXiv.2406.11903.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2024. Multilegalpile: A 689gb multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15077–15094. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint*, arXiv.2408.10343.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11563–11579. Association for Computational Linguistics.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Jingyun Sun, Zhongze Luo, and Yang Li. 2025. A compliance checking framework based on retrieval augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2603–2615, Abu Dhabi, UAE. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8818–8847. Association for Computational Linguistics.

Thomson Reuters Regulatory Intelligence. 2023. Cost of compliance report 2023. Technical report, Thomson Reuters. 14th annual edition. https://legal.thomsonreuters.com/en/insights/reports/cost-of-compliance-2023.

Santosh T.y.s.s, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, and Xiaomo Liu. 2025. CoCoLex: Confidence-guided copy-based decoding for grounded legal text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19002–19018, Vienna, Austria. Association for Computational Linguistics.

Santosh T.y.s.s and Elvin Quero Hernandez. 2025. LexKeyPlan: Planning with keyphrases and retrieval augmentation for legal text generation: A case study on European court of human rights cases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–436, Vienna, Austria. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint*, arXiv.2310.07521.

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

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint*, arXiv.2306.05443.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint*, arXiv.2306.06031.

Dániel Zagyva, Emmanouil Stergiadis, Laurens Van Der Maas, Aleksandra Dokic, Eran Fainman, Ilya Gusev, and Moran Beladev. 2025. Speed without sacrifice: Fine-tuning language models with medusa and knowledge distillation in travel applications. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 684–692. Association for Computational Linguistics.

A GRC Platform Comparison

Capability	Ascent	CUBE	OneSumX	Ours
Reg. change tracking	✓	✓	✓	✓
Automated obligation extr.	–	Partial	–	✓
Multi-framework gap analysis	–	–	–	✓
Grounding verification	–	–	–	✓
Sub-second latency	N/A	N/A	N/A	✓†

Table 5: Qualitative comparison with commercial GRC platforms. Based on published product documentation (Jain et al., 2025); checkmarks represent vendor-reported capabilities, not independently verified performance. N/A = manual workflow. †p50; p99 is 1,082 ms. The commercial GRC market is changing quickly, with Thomson Reuters, Wolters Kluwer, and CUBE recently announcing LLM-based capabilities.

B Algorithm and Notation

Symbol	Description
α	Dense/sparse retrieval weight (default 0.7)
β	KG re-ranking weight (default 0.3)
δ	Gap detection threshold (default 0.6; deploy 0.45)
τ	Grounding confidence threshold (default 0.85)
γ	KD/SFT balance weight (default 0.5)
$\lambda_{1,2,3}$	Multi-task loss weights (0.4, 0.3, 0.3)

Table 6: Summary of key hyperparameters.

C Dataset Details

RegObligation. The 1,847 regulatory sentences span SEC (712), MiFID II (614), and Basel III (521). Difficulty varies: 62% involve single cross-references, 27% require resolving 2 cross-references, and 11% require 3+. We plan to release REGOBLIGATION upon publication.

GapBench. 423 obligation–policy pairs from anonymized compliance reviews at a single financial institution. Framework distribution: SEC (178), MiFID II (143), Basel III (102). GAPBENCH cannot be released in its current form due to institutional data use agreements, but we are working on an anonymized version with synthetic policy substitutions.

Algorithm 1 ComplianceNLP End-to-End Pipeline

Require: Regulatory document D, internal policies $\mathcal{P}$, RKG $\mathcal{G}$, thresholds τ, δ
Ensure: Compliance gap report R
Stage 1: Obligation Extraction
$\{o_1, \ldots, o_n\} \leftarrow \text{MultiTask}_{\mathcal{E}}(D)$ {NER + Deontic + XRef}
Resolve cross-references via $\mathcal{G}$ lookup
Stage 2: Retrieval & Grounding
for each obligation o_j **do**
  $\mathcal{C}_j \leftarrow \text{Retrieve}(o_j, \alpha)$ {Hybrid dense+BM25}
  $\mathcal{C}_j^* \leftarrow \text{KGRerank}(\mathcal{C}_j, \mathcal{G}, \beta)$
end for
Stage 3: Gap Analysis
for each obligation o_j **do**
  $p^* \leftarrow \arg\max_{p_k \in \mathcal{P}} a(o_j, p_k)$
  if $a(o_j, p^*) < \delta$ **then**
    $\text{gap}_j \leftarrow \text{Generate}(o_j, p^*, \mathcal{C}_j^*)$
    Verify grounding via MiniCheck (τ)
    Flag low-confidence outputs for human review
  end if
end for
$R \leftarrow \text{CompileReport}(\{\text{gap}_j\})$
return R

D RKG Schema

Five node types: PROVISION (a regulatory article, section, or clause), ENTITY (a regulated entity type), OBLIGATION (a structured obligation with modality, action, condition), THRESHOLD (a quantitative threshold, e.g., Basel III CET1 $\geq$ 4.5%), and ENFORCEMENT (a documented enforcement action with penalty amount, date, and entity).

Five typed edges: AMENDS, SUPERSEDES, CROSSREFERENCES (provision $\rightarrow$ provision), IMPLEMENTS (provision $\rightarrow$ obligation), APPLIESTO (obligation $\rightarrow$ entity). Statistics: 12,847 provision nodes (SEC: 4,932; MiFID II: 4,218; Basel III: 3,697), 1,247 entity nodes, 8,431 obligation nodes, 612 threshold nodes, 847 enforcement nodes, 34,219 edges.

The primary source of missed edges is implicit cross-references lacking explicit citation markers (e.g., "consistent with the provisions governing capital requirements" without specifying the article), accounting for 68% of recall gaps. Novel cross-reference patterns flagged by the linker's low-confidence scores are queued for human review.

Update mechanics. The RKG is maintained through a three-stage update cycle. (1) *Trigger-*

ing: nightly synchronisation pulls regulatory feeds (SEC EDGAR XML, EUR-Lex HTML, BIS PDF); out-of-cycle updates fire when (a) a manual trigger is issued by a compliance analyst, or (b) the monitoring dashboard detects a confidence-score anomaly (rolling-mean drop $>2\sigma$) on incoming documents, indicating distributional shift relative to recent ingestions. (2) *Incremental incorporation*: parsed provisions are inserted as new PROVISION nodes; cross-reference edges are added by re-running the span-pair linker over the new provision text against the existing KG, scoped to candidate targets within a 2-hop neighbourhood of provisions sharing the same article-level ancestor. Embedding updates are restricted to this 2-hop neighbourhood rather than recomputed globally. The FAISS vector store is rebuilt nightly; between rebuilds, new chunks are indexed via approximate nearest-neighbour insertion. (3) *Consistency control*: provisions added since the last full synchronisation carry a `pending_validation` flag that prevents them from contributing to KG re-ranking scores until the next full rebuild validates their edges; incoming queries that touch flagged provisions transparently fall back to embedding-only retrieval. The maximum blind-spot window is therefore 18 hours; during that window, ablation evidence (Table 2, "w/o KG reranking", -4.6 gap F1) bounds the cost of the fallback. For routine monitoring this fallback is acceptable, since institutional implementation windows for new regulations typically exceed 30 days.

E MiniCheck Validation Details

Metric	Value
Sample size	200
Human IAA (κ)	0.87
MiniCheck–Human agreement	91.8%
MiniCheck precision (vs. human)	93.2%
MiniCheck recall (vs. human)	90.4%
Pearson r (score vs. human)	0.83
Disagreement analysis (n=16 disagreements)	
Provision confusion (e.g., Art. 25(1) vs. 25(2))	62%
Subtle modality errors (should vs. shall)	25%
Incomplete citation	13%

Table 7: MiniCheck validation against human judgments on 200 regulatory text samples.

F False Negative Analysis

We manually examined the 17 missed FULL GAP cases and 23 missed PARTIAL GAP cases. Primary failure modes: (i) *implicit obligations* (35%), where requirements are implied rather than explicitly stated with deontic modals; (ii) *multi-hop cross-references* (29%), where the obligation depends on resolving 3+ cross-references; and (iii) *jurisdiction-specific nuance* (21%), where the gap exists only under a specific national implementation.

Per-class significance for ablations. Given limited support for FULL GAP (n=85): removing KG re-ranking drops Full Gap F1 from 81.5 to 76.1 ($\Delta=-5.4$, CI: $[-9.1, -1.8]$, p=0.007). A power analysis indicates $\sim$200 Full Gap examples needed to detect a 3-point F1 difference at 80% power.

G Additive Component Analyses

We provide two complementary additive analyses starting from different base models to confirm that component attributions are robust to the choice of base. Table 8 starts from GPT-4o+RAG; Table 9 starts from LLaMA-3-8B+RAG. Marginal contributions (Δ) for KG re-ranking, multi-task extraction, and MiniCheck verification are reported below each configuration; the rankings are consistent across both base models, with KG re-ranking providing the largest gap-detection gain in both cases.

Configuration	NER	Deon.	Gap	EM	F1
GPT-4o + RAG (base)	88.6	90.5	84.2	48.1	66.8
+ KG re-ranking	89.8	91.1	86.4	49.7	68.3
+ Multi-task	90.9	92.4	87.1	52.0	71.0
+ MiniCheck	**91.3**	**92.7**	**87.7**	**52.8**	**71.9**
Marginal Δ:					
KG re-ranking	+1.2	+0.6	+2.2	+1.6	+1.5
Multi-task extraction	+1.1	+1.3	+0.7	+2.3	+2.7
MiniCheck verification	+0.4	+0.3	+0.6	+0.8	+0.9

Table 8: Additive analysis from GPT-4o+RAG base. KG re-ranking provides the largest gap detection gain (+2.2); multi-task extraction contributes most to QA (+2.7 F1). MiniCheck's primary impact is on grounding accuracy (86.7%$\rightarrow$94.2%), not captured by F1.

H Additional Experimental Results

This appendix collects supplementary experimental results referenced from §6: per-system standard deviations across three seeds (Table 10), per-framework performance breakdown (Table 11), and end-to-end error propagation from extraction to gap detection (Table 12).

Per-system standard deviations.

Configuration	NER	Deon.	Gap	EM	F1
LLaMA-3-8B+RAG (base)	87.9	89.8	83.5	47.4	65.9
+ KG re-ranking	89.4	90.6	86.0	49.2	67.8
+ Multi-task	90.6	92.1	86.8	51.7	70.5
+ MiniCheck	**91.3**	**92.7**	**87.7**	**52.8**	**71.9**
Marginal Δ*:*					
KG re-ranking	+1.5	+0.8	+2.5	+1.8	+1.9
Multi-task extraction	+1.2	+1.5	+0.8	+2.5	+2.7
MiniCheck verification	+0.7	+0.6	+0.9	+1.1	+1.4

Table 9: Complementary additive analysis from LLaMA-3-8B+RAG base. Component rankings are consistent with Table 8: KG re-ranking provides the largest gap detection gain (+2.5 here vs. +2.2), multi-task extraction contributes most to QA (+2.7 in both), confirming robustness to base model choice.

System	NER	Deon.	Gap	QA F1
GPT-4o (5-shot)	$\pm$0.2	$\pm$0.1	$\pm$0.3	$\pm$0.2
GPT-4o + RAG	$\pm$0.2	$\pm$0.2	$\pm$0.3	$\pm$0.3
LLaMA-3-8B + RAG	$\pm$0.3	$\pm$0.2	$\pm$0.4	$\pm$0.3
LLaMA-3-70B*	$\pm$0.3	$\pm$0.2	$\pm$0.4	$\pm$0.3
COMPLIANCENLP	$\pm$0.2	$\pm$0.1	$\pm$0.3	$\pm$0.2

Table 10: Per-system standard deviations (3 seeds) for main results in Table 2.

Per-framework breakdown.

Framework	# Sent.	NER	Deon.	XRef	Ground.
SEC	712	93.1	94.2	89.7	95.8
MiFID II	614	91.4	92.8	85.3	94.1
Basel III	521	88.4	90.6	81.2	92.4
All	1,847	91.3	92.7	85.9	94.2

Table 11: Per-framework results on REGOBLIGATION. SEC achieves highest performance due to standardized filing formats. Basel III is most challenging: XRef F1 drops to 81.2 due to nested multi-level cross-references spanning accord revisions. In deployment, Basel III documents receive mandatory human review.

Grounding accuracy by generation length. 96.1% for outputs <50 tokens, 93.8% for 50–150 tokens, 89.4% for >150 tokens.

Error analysis. Manual examination of 100 obligation extraction errors: boundary errors in long nested entities (34%), cross-reference resolution failures for multi-document references (28%), ambiguous deontic modality with hedged phrasing (19%), entity type confusion (11%), and annotation disagreements (8%).

Operational impact. The 83.4 end-to-end F1 translates to approximately 2.1 missed obligations per 100-page regulatory document and 1.3 false gap alerts per day. At deployment threshold (δ=0.45), false alerts increase to 3.7/day while missed obligations drop to 0.7/document.

End-to-end error propagation.

Gap Input	Gap F1	Δ
Gold obligations	87.7	–
Predicted obligations	83.4	−4.3
Propagation source breakdown		
NER boundary errors	–	−2.9
Cross-reference failures	–	−1.0
Deontic misclassification	–	−0.4

Table 12: End-to-end error propagation from obligation extraction to gap detection (150 documents).

I Hyperparameter Sensitivity

Table 13 reports sensitivity to the four key hyperparameters of the retrieval and gap-detection pipeline. Results are robust across reasonable ranges, with all four parameters varying gap detection F1 by less than $\pm$2 points.

J Annotation Guidelines and Entity Distribution

Annotation process. Three compliance experts annotated REGOBLIGATION using a custom web-based interface. Annotators were provided with a 12-page guideline document specifying: (a) entity type definitions with 2–3 examples each, (b) boundary rules for nested and discontinuous entities, (c) deontic modality decision tree for ambiguous cases, and (d) cross-reference resolution rules. Training consisted of a 2-hour session followed by 50 practice sentences with adjudication. Inter-annotator agreement was κ=0.84 (Fleiss). Disagreements were resolved by majority vote, with a senior analyst adjudicating the 3.2% of three-way disagreements.

NER entity type distribution. The 23 NER entity types and their relative frequencies in REGOBLIGATION: REGULATORY_BODY (13.8%), REPORTING_ENTITY (11.5%), EFFECTIVE_DATE (9.0%), THRESHOLD_VALUE (8.4%), FINANCIAL_INSTRUMENT (7.2%), OBLIGATION_ACTION (6.7%), COMPLIANCE_PERIOD (5.9%), JURISDICTION (5.6%), PENALTY_AMOUNT (4.2%), RISK_CATEGORY (3.8%), CAPITAL_REQUIREMENT (3.4%), DISCLOSURE_ITEM (3.1%), FILING_TYPE (2.7%), COUNTERPARTY (2.3%), SU-

Parameter	Range	Default	F1 Range	Δ
α (retrieval)	[0.5, 0.8]	0.7	90.5–91.3	±0.8
δ (gap threshold)	[0.4, 0.7]	0.6	85.9–87.7	±1.8
k (top-k passages)	[3, 7]	5	90.7–91.3	±0.6
β (KG rerank)	[0.1, 0.5]	0.3	90.1–91.3	±1.2

Table 13: Sensitivity analysis for key hyperparameters. Results are robust across reasonable ranges.

PERVISORY_AUTHORITY (2.0%), MARKET_TYPE (1.8%), TRANSACTION_TYPE (1.7%), GOVERNANCE_ROLE (1.5%), AUDIT_REQUIREMENT (1.3%), REPORTING_FREQUENCY (1.2%), LEGAL_REFERENCE (1.1%), CROSS_BORDER_PROVISION (1.1%), EXEMPTION_CLAUSE (0.8%).

K Deployment Details

Deployment timeline. The system has progressed through four phases: **Phase 1** (completed, 6 months): offline development, architecture validation, and benchmark evaluation. **Phase 2** (current, 4 months in): integration with the institution's GRC platform; parallel operation alongside manual processes with 12 compliance analysts across 3 regulatory teams. **Phase 3** (planned, Q3 2026): shadow deployment on live regulatory feeds with graduated autonomy for low-severity findings. **Phase 4** (target Q1 2027): full production deployment with human review limited to Critical/Major findings.

Integration architecture. COMPLIANCENLP integrates via: (i) regulatory feed ingestion from SEC EDGAR and EUR-Lex RSS feeds, synchronized nightly into the RKG (Neo4j) and vector store (FAISS); (ii) structured JSON API outputs conforming to the institution's internal schema; and (iii) case management integration for remediation tracking. The system runs on AWS with auto-scaling GPU instances (A100 80GB).

Monthly performance stability. Monthly production metrics: Month 1: precision 88.2%, recall 94.1%, override rate 11.4%; Month 2: 89.9%, 95.3%, 9.1%; Month 3: 91.8%, 96.8%, 7.2%; Month 4: 92.4%, 97.3%, 6.1%. We acknowledge that improving trends may partially reflect changes in regulatory update difficulty distribution or analyst adaptation.

User study. During Phase 2, 12 compliance analysts processed 96 regulatory updates (8 per analyst) both manually and with system assistance (counterbalanced order, unblinded). System-assisted review reduced average processing time from 47 to 12 minutes ($3.9\times$, $p<0.01$, paired t-test). In 89/96 updates (92.7%), system-assisted determinations matched manual. Processing time stabilized at 15 min/update ($3.1\times$) after novelty attenuation.

Severity	Prod. P	Prod. R	Prod. F1	GapBench F1
Compliant	94.8	96.1	95.4	93.2
Partial Gap	88.3	91.7	90.0	83.7
Full Gap	86.1	93.4	89.6	81.5
Macro avg.	89.7	93.7	91.7	86.1

Table 14: Production-derived accuracy by severity (δ=0.45) vs. GapBench (δ=0.6). Production outperforms GapBench because real-world gaps skew toward more explicit obligation language.

Production-derived accuracy by severity.

Missed gap analysis. The 47 missed gaps were predominantly in Basel III amendment chains (23/47) and implicit MiFID II obligations (14/47). By severity: 3 Critical (6.4%), 11 Major (23.4%), 19 Moderate (40.4%), 14 Minor (29.8%). The system identified 31 potential gaps the manual process had not flagged; 24 were subsequently confirmed.

L Production Recall Estimation Caveats

The 96.0% production recall is an *estimate* because the true total number of compliance gaps is unknowable. Both the system and manual process may share blind spots. The denominator is estimated as gaps caught by the system plus gaps missed by the system but caught by manual review. Under sensitivity analysis: if the manual process has 95% recall, estimated system recall adjusts to 95.8%; at 90% manual recall, 95.6%. The adjustment is small because the probability of both processes missing the same gap is low under approximate independence.

M Cost Analysis

N Cross-Institutional Generalization

Taxonomy mapping. The primary adaptation challenge is on the policy alignment side: the European partner uses a different taxonomy (organized by business function rather than regulation), different naming conventions, and different granularity, requiring re-training f_{type} but not affecting upstream extraction or KG components.

Phase	Infra.	Analyst	Total	Δ
Manual (baseline)	\$0	\$45K	\$45K	–
Phase 2 (current)	\$11.2K	\$37K	\$48.2K	+7%
Phase 3 (projected)	\$11.2K	\$22K	\$33.2K	−26%
Phase 4 (projected)	\$11.2K	\$9K	\$20.2K	−55%

Table 15: Monthly cost breakdown. Phase 2 is higher than manual because all outputs require review. Infrastructure: two A100 instances (∼\$8.1K/month) plus Neo4j and FAISS hosting. At 2,400 docs/month, cost is \$4.67/doc.

Controlled perturbation experiment. Replacing our institution's policies with synthetically perturbed policies on 150 GapBench pairs: extraction was unaffected; gap detection F1 dropped by 3.2 (moderate perturbation) and 6.8 (aggressive perturbation), supporting our estimate that ∼70% of difficulty is regulation-specific and ∼30% institution-specific.

Per-class cross-institutional breakdown. On 87 pairs from the European partner (without / with 50-example adaptation): COMPLIANT: 89.2 / 91.4 F1 (n=41); PARTIAL GAP: 79.8 / 82.1 F1 (n=28); FULL GAP: 78.4 / 82.7 F1 (n=18). Adaptation primarily improved the alignment module (+3.1 F1 on Partial/Full Gap combined).

Forward-looking risk. As the system transitions to graduated autonomy, automation bias risk may re-emerge. We plan to maintain weekly calibration and introduce random audit sampling (5% of low-severity outputs).

O Additional Lessons Learned

This appendix supplements the five key lessons in §7 with additional deployment insights. Two of these expand lessons 4 (GRC integration) and 5 (organisational trust-building) summarised in the main text, and two further insights cover what did not work and the monitoring of novel regulatory structures. We hope they benefit practitioners building NLP systems for regulated domains.

What didn't work: end-to-end fine-tuning. Full fine-tuning of LLaMA-3-70B on our regulatory corpus degraded general reasoning capabilities needed for gap analysis (gap F1: 86.3→81.7) and cost ∼\$2,800/run on 8×A100. The degradation was most pronounced on gap descriptions requiring multi-step reasoning across provisions: the fine-tuned model frequently "shortcut" to pattern-matched outputs rather than synthesizing cross-reference chains. We hypothesize this reflects catastrophic forgetting of general reasoning capabilities during domain adaptation, consistent with broader findings on aggressive domain-specific fine-tuning of large language models. Knowledge distillation to an 8B model proved far more practical: it preserved the teacher's reasoning structure while achieving the latency targets, and each distillation run cost ∼\$420, an order of magnitude cheaper than full fine-tuning with superior downstream performance.

GRC integration is harder than model development. Integrating COMPLIANCENLP into the institution's existing Governance, Risk, and Compliance (GRC) platform consumed approximately 3 months of engineering effort, comparable to the entire model development cycle. The primary challenge was mapping our system's output taxonomy to a 15-year-old internal compliance taxonomy with 847 categories organized by regulatory source, while the GRC platform expected categorization by business function. This required building a bidirectional mapping layer with 94.1% automated classification accuracy and a manual review queue for ambiguous cases (5.9% of outputs). Additional integration friction included: (1) API rate limits on the legacy GRC system requiring batched writes with retry logic; (2) audit trail requirements mandating full provenance chains from source regulation through extraction to gap classification, adding ∼340 ms per document; and (3) role-based access control alignment across three separate authentication systems. We recommend that future deployments budget at least 40% of total project timeline for GRC integration and begin taxonomy mapping in parallel with model development rather than sequentially.

Organizational adoption requires staged trust-building. Deploying an automated compliance system into a risk-averse organizational culture required deliberate trust calibration. In Month 1, system-assisted processing time was 22 min/update, only 2.1× faster than manual, because analysts independently re-verified the majority of system outputs (78% full re-verification rate). By Month 4, processing time had dropped to 12 min/update (3.9×) with a re-verification rate of 23%, concentrated on high-severity and cross-framework outputs. The single most effective trust-building intervention was publishing weekly "system vs. manual" accuracy digests that transparently reported both

system successes and failures. This was more effective than training sessions or documentation alone, because analysts could track cumulative evidence rather than relying on point demonstrations. We also observed an initial automation bias risk: 34% of analysts accepted system outputs without independent verification in Weeks 1–2, which we mitigated through mandatory confidence-score display and weekly calibration exercises where analysts reviewed deliberately seeded errors. By Month 3, uncritical acceptance dropped to 12%. We recommend that deployments in high-stakes domains plan for a 2–3 month trust calibration period and invest in transparent performance reporting from day one.

Novel regulatory structures require active monitoring. During the parallel run, a Basel III amendment introduced obligation structures not well-represented in training data: specifically, nested conditional obligations with temporal triggers spanning multiple articles. Confidence scores on affected documents dropped sharply (mean 0.62 vs. 0.84 overall), which our monitoring dashboard flagged within 6 hours. This enabled targeted retraining within 48 hours, restoring performance to baseline levels. Over 4 months, we detected 3 distributional shift events through a combination of confidence score monitoring (threshold: 2σ below rolling mean), embedding drift detection on incoming documents, and analyst feedback signals from the review interface. These events required 2 retraining cycles (each $\sim$8 hours on $2\times$A100) and 1 parser update for a newly introduced SEC cross-reference format. Ongoing steady-state maintenance requires $\sim$4–6 hours/week for KG review queue processing and 1 model update per 6–8 weeks. Deploying NLP systems in evolving regulatory environments is not a one-time effort: the monitoring and maintenance infrastructure is as critical as the initial model, and teams should plan for dedicated MLOps capacity post-deployment.

P Additional Limitations

Beyond the seven limitations discussed in the main text, we identify additional concerns that qualify our results and inform future work.

User study design. Our user study (12 analysts, 96 updates) is modest in scale and unblinded, introducing two methodological concerns. First, the sample size limits statistical power for subgroup analyses: we cannot reliably estimate whether efficiency gains differ by analyst seniority, regulatory framework, or gap severity. A post-hoc power analysis indicates that detecting a 20% difference in processing time between junior and senior analysts would require $\sim$40 participants per group. Second, the unblinded design means analysts were aware they were using COMPLIANCENLP, which may inflate efficiency gains through Hawthorne effects or deflate them through increased scrutiny. While the post-study stabilization at $3.1\times$ (vs. $3.9\times$ during the controlled period) partially addresses novelty effects, a blinded study, where analysts receive system outputs embedded within a standard review interface without attribution, would provide stronger causal evidence. We were unable to implement blinding due to institutional constraints on modifying the existing compliance workflow during the parallel run.

Grounding verification limitations. MiniCheck-DeBERTa achieves r=0.83 correlation with human grounding judgments on our validation set, but this aggregate figure masks performance variation across obligation complexity. On simple single-provision obligations, agreement reaches 95.4%; however, on obligations requiring 3+ cross-references, grounding accuracy drops to 84.6%, and on nested conditional obligations spanning multiple articles, it further degrades to an estimated $\sim$79% (based on a smaller sample of 38 instances). The primary failure mode is conflating related but distinct provisions; for example, treating a delegated regulation's implementing measure as equivalent to its parent directive article. MiniCheck also operates at the sentence level and may miss sub-article-level distinctions where a provision's specific clause (e.g., Article 25(2)(a) vs. 25(2)(b)) materially affects the compliance assessment. We mitigate this through the mandatory analyst review of all High and Critical severity findings, but acknowledge that a more granular grounding model calibrated at the clause level would strengthen the system's autonomous reliability.

Latency tail and throughput constraints. While median end-to-end latency (4.2 min/doc) and p50 inference latency (659 ms) meet operational requirements, the p99 inference latency of 1,082 ms exceeds our sub-second target. Tail latency events correlate with three factors: (1) documents triggering $>$5 KG traversal hops during re-ranking (accounting for 41% of p99 events), which arise from

deeply nested cross-reference chains in Basel III amendment structures; (2) Medusa draft rejection cascades, where >3 consecutive rejections force sequential decoding fallback (29% of p99 events); and (3) FAISS index contention under concurrent queries during peak processing windows (30% of p99 events). The first factor is addressable through KG traversal depth limits with graceful degradation, though this risks missing deep cross-references. The second is partially mitigated by our domain-specific Medusa heads but remains inherent to speculative decoding on atypical regulatory language. The third is an infrastructure scaling issue resolvable with index sharding. For our deployment context, where documents are processed in batch rather than interactively, tail latency has minimal operational impact, though it would be a concern for real-time advisory use cases.

Knowledge graph currency and coverage. The KG operates on a nightly update cycle synchronized with regulatory feeds, introducing an 18-hour maximum blind spot for newly published regulations. During this window, the system relies on embedding-based retrieval alone (without KG re-ranking), which our ablations show reduces gap detection F1 by 4.6 points. For routine regulatory updates, this delay is operationally acceptable, since most institutions allow 30+ day implementation windows. However, for emergency regulatory actions (e.g., emergency SEC orders or temporary supervisory measures), the blind spot could result in delayed gap identification. We are investigating streaming KG updates through an incremental graph ingestion pipeline, though this introduces consistency challenges when new provisions create edges to not-yet-indexed related provisions. Our KG also covers 12,847 provisions across three frameworks; provisions from adjacent but uncovered frameworks (e.g., Dodd-Frank Title VII, EMIR) that cross-reference covered provisions are not represented, potentially causing incomplete cross-reference resolution at framework boundaries.

Q Example System Output

We present three end-to-end examples, one per regulatory framework, illustrating the full pipeline from regulatory input through obligation extraction, KG-augmented retrieval, and gap analysis. Each example shows the structured obligation representation ⟨*entity*, *action*, *modality*, *condition*, *source_provision*⟩ defined in §3.3, the KG retrieval and re-ranking trace (§3.1), and the final gap classification with grounding verification (§6.1). Confidence scores reflect the deployment threshold (δ=0.45).

Q.1 Example 1: MiFID II Client Suitability (Major Gap)

Input. MiFID II update requiring investment firms to assess client suitability (Article 25(2) of Directive 2014/65/EU).

Multi-task extraction output. The LEGAL-BERT extraction module identifies one primary obligation with two cross-references:

> **Entity:** Investment firm [NER: `REGULATED_ENTITY`, conf: 0.97]
> **Action:** Obtain information regarding client knowledge, experience, financial situation, and investment objectives
> **Modality:** OBLIGATION ("shall") [DEONTIC conf: 0.94]
> **Condition:** When providing investment advice or portfolio management
> **Source:** MiFID II Art. 25(2)
> **Cross-references resolved:** Delegated Regulation (EU) 2017/565 Art. 54–56 (suitability assessment procedures), Art. 25(3) (appropriateness for non-advised services) [XREF conf: 0.91, 0.88]

KG retrieval trace. Hybrid retrieval returns 5 candidate passages (α=0.7). KG re-ranking promotes two passages from Delegated Regulation Art. 54–56 (graph distance: 1 hop from Art. 25(2)) that were ranked 4th and 5th by dense retrieval alone, displacing two tangentially related MiFID I provisions. Re-ranked top-3: Art. 25(2) source text (s_{KG}=0.94), Delegated Regulation Art. 54 (s_{KG}=0.89), Art. 55 (s_{KG}=0.86).

Compliance gap detected.

> **Classification:** PARTIAL GAP [alignment score: 0.52, below δ=0.6]
> **Severity:** MAJOR
> **Matched policy:** Internal Policy §4.3 "Client Onboarding"
> **Gap description:** Current policy requires collection of client financial situation and investment objectives at onboarding but does not mandate re-assessment of knowledge and experience at each advisory interaction, as required by Art. 25(2). The policy also lacks reference to the detailed suitability assessment procedures specified in Delegated Regulation Art. 54–56.
> **Recommended action:** Update §4.3 to include periodic suitability reassessment per Delegated Regulation Art. 54 and add cross-reference to Art. 55–56 assessment criteria.
> **Grounding:** 3/3 sentences verified against source provisions (MiniCheck confidence: 0.94).
> **Analyst action:** Confirmed. Gap escalated to policy review committee.

Q.2 Example 2: SEC Regulation S-K Disclosure (Compliant)

Input. SEC amendment to Regulation S-K Item 402, requiring disclosure of the relationship between executive compensation and financial performance (17 CFR §229.402, effective 2023).

Multi-task extraction output.

> **Entity:** Registrant [NER: `REPORTING_ENTITY`, conf: 0.96]
> **Action:** Provide a clear description of the relationship between executive compensation actually paid and cumulative TSR, net income, and a company-selected measure
> **Modality:** OBLIGATION ("must") [DEONTIC conf: 0.97]
> **Condition:** In any proxy or information statement for which disclosure under Item 402 is required
> **Source:** Reg. S-K Item 402(v)
> **Cross-references resolved:** Exchange Act Rule 14a-3 (proxy solicitation requirements), Item 402(c) (summary compensation table) [XREF conf: 0.93, 0.90]

KG retrieval trace. The SEC EDGAR XML parser provides well-structured source provisions, enabling high-confidence retrieval. Hybrid retrieval returns Item 402(v) source text (s_{KG}=0.96), Item 402(c) summary compensation table requirements (s_{KG}=0.91, graph distance: 1 hop), and Exchange Act Rule 14a-3 (s_{KG}=0.87, 2 hops). KG re-ranking has minimal effect on SEC texts due to the standardized filing structure. This is consistent with our observation that SEC texts yield the highest NER F1 (93.1) among the three frameworks (§6).

Compliance assessment.

> **Classification:** COMPLIANT [alignment score: 0.78, above δ=0.6]
> **Severity:** N/A
> **Matched policy:** Internal Policy §7.1 "Executive Compensation Disclosure" and §7.4 "Pay-vs-Performance Tabular Disclosure"
> **Assessment:** Existing policy §7.1 mandates disclosure of the relationship between executive compensation and financial performance including TSR and net income, and §7.4 requires tabular presentation of pay-versus-performance data. Both cross-referenced SEC requirements are addressed in current policy language.
> **Grounding:** 2/2 sentences verified (MiniCheck confidence: 0.96).
> **Analyst action:** Confirmed compliant. No action required.

Q.3 Example 3: Basel III Liquidity Coverage Ratio (Critical Gap)

Input. Basel III amendment to the Liquidity Coverage Ratio (LCR) framework, introducing revised High-Quality Liquid Asset (HQLA) eligibility criteria and jurisdictional treatment of central bank reserves (BCBS d424, paragraphs 49–52, cross-referencing BCBS d295 and CRR Article 412).

Multi-task extraction output. This example illustrates a complex multi-hop extraction chain spanning three documents:

> **Entity:** Internationally active bank [NER: `REGULATED_ENTITY`, conf: 0.93]
> **Action:** Include only eligible central bank reserves in Level 1 HQLA, subject to operational requirements and jurisdictional supervisory approval for drawdown during stress
> **Modality:** OBLIGATION ("shall") [DEONTIC conf: 0.91]
> **Condition:** When calculating LCR numerator under normal and stressed conditions
> **Source:** BCBS d424 ¶50
> **Cross-references resolved:** BCBS d295 ¶28–31 (original HQLA operational requirements), CRR Art. 412(1) (EU-specific LCR implementation), BCBS d424 ¶52 (jurisdictional supervisory discretion on reserve drawdown) [XREF conf: 0.89, 0.84, 0.92]

Note the lower confidence on the CRR Art. 412(1) cross-reference (0.84), reflecting the cross-framework boundary between Basel Committee standards and EU implementation, a known challenge area where our KG has incomplete coverage of adjacent frameworks (Appendix P).

KG retrieval trace. This is the most challenging retrieval scenario in our examples. The three-document cross-reference chain requires 3 KG hops: BCBS d424 ¶50 → BCBS d295 ¶28 → CRR Art. 412(1). Dense retrieval alone ranked BCBS d295 ¶28 at position 7 (below top-k=5); KG re-ranking promoted it to position 2 (s_{KG}=0.82). CRR Art. 412(1) was not in the initial retrieval set; it was added through KG traversal at hop 3, with s_{KG}=0.74. Without KG re-ranking, this obligation would lack the CRR context entirely, illustrating why the KG contributes +4.6 gap F1 in ablations. This example falls in the category of Basel III amendment chains that account for 23/47 (49%) of missed gaps in production (§7).

Compliance gap detected.

> **Classification:** FULL GAP [alignment score: 0.31, below δ=0.45]
> **Severity:** CRITICAL
> **Matched policy:** Internal Policy §12.2 "Liquidity Buffer Composition" (best match)
> **Gap description:** Current policy defines HQLA eligibility based on the original BCBS d295 criteria and does not incorporate the revised eligibility criteria from d424, specifically: (1) the updated

operational requirements for central bank reserve inclusion in ¶49–50; (2) the jurisdictional supervisory approval requirement for reserve drawdown during stress in ¶52; and (3) the interaction with CRR Art. 412(1) for EU-supervised entities. The policy references BCBS d295 but not the superseding d424 provisions.
Recommended action: Immediate revision of §12.2 to incorporate BCBS d424 HQLA criteria. Engage with jurisdictional supervisor regarding reserve drawdown approval. Update CRR mapping table for EU entity compliance.
Grounding: 4/5 sentences verified against source provisions (MiniCheck confidence: 0.87). Sentence 3 received a borderline score (0.71) due to the multi-hop CRR cross-reference, flagged for mandatory analyst review per our grounding policy for 3+ cross-reference obligations.
Analyst action: Confirmed Critical gap. Escalated to Head of Treasury within 24 hours per internal SLA. CRR cross-reference manually verified and confirmed.

Q.4 Summary of Examples

These three examples illustrate the system's behavior across the difficulty spectrum: a routine SEC compliance confirmation where KG re-ranking has minimal impact, a MiFID II partial gap requiring cross-reference to delegated regulations, and a complex Basel III multi-hop scenario where KG traversal is essential. The Basel III example also demonstrates the system's behavior at its performance boundaries: lower extraction confidence, borderline grounding scores, and mandatory analyst escalation. This is consistent with our finding that Basel III documents receive mandatory human review in deployment (§6).