Recursive Norm Refinement for Modern Slavery Risk Reasoning in Social LLM Systems

Anonymous Author(s)

Anonymous Institution

Abstract. Social computing systems increasingly mediate recruitment, platform work, migration assistance, welfare triage, supply-chain monitoring, and public reporting about labour exploitation. Large language model (LLM) agents can help structure these tasks, but they also create a distinctive legal risk: fluent outputs can collapse binding treaties, soft-law guidance, regional instruments, and scholarly commentary into undifferentiated claims about "modern slavery." This paper proposes *Recursive Norm Refinement* (RNR), a verifier-governed architecture for LLM agents that reason about modern slavery risk under international human rights law. RNR decomposes generated answers into atomic legal claims, verifies each claim against a source hierarchy, checks actor, jurisdiction, temporal scope, deontic modality, exception entailment, and evidentiary confidence, and permits memory promotion only for verified claims. We instantiate the architecture for forced labour, slavery, servitude, trafficking, forced population transfer, and business human-rights due diligence. We report three diagnostic experiments: an adversarial claim-verification benchmark derived from primary legal instruments; a recursive-memory simulation measuring contamination from unsupported legal claims; and a rights-aware triage exercise for AI-mediated labour-risk scenarios. The experiments are not a leaderboard for any model; they test whether the proposed controls detect source inflation, modality drift, jurisdictional overreach, and over-generalized risk claims. Results show that full RNR substantially reduces false acceptance and memory contamination compared with retrieval-only and source-only ablations, while making remaining failure modes explicit. The result is a verifier-governed pipeline that can be instantiated, tested, and extended for legally grounded social LLM systems in high-stakes modern slavery contexts.

Keywords: Modern Slavery · Forced Labour · Legal LLMs · Social Computing · Human-Rights Due Diligence · Benchmark Design

1 Introduction

Modern slavery is not a metaphor for generic unfairness. In legal and policy practice it covers partly overlapping categories: slavery, servitude, forced labour, trafficking, debt bondage, forced marriage, child exploitation, and related coercive labour control. The ILO, Walk Free, and IOM estimate about 50 million

persons in modern slavery in 2021, including about 28 million in forced labour and 22 million in forced marriage [1]. The legal framework is dense. ILO Convention No. 29 defines forced labour by work or service, menace of penalty, and absence of voluntary offer; ILO Convention No. 105 prohibits specified uses such as political coercion, labour discipline, strike punishment, and discrimination [2, 3]. Slavery, trafficking, regional human-rights instruments, humanitarian law, and international criminal law add different actors, thresholds, contexts, and consequences [4–6, 8, 10, 11, 7, 12, 13].

Social computing changes how these norms are encountered. Labour platforms infer reliability and allocate jobs. Recruitment systems sort workers before any human interview. Supply-chain analytics infer forced-labour risk from shipping records, audit text, media reports, and worker communications. Social media analysis tracks trafficking narratives and public reports. Humanitarian and migration systems classify vulnerability. Generative systems are now inserted into these workflows as summarizers, triage assistants, risk classifiers, draft auditors, and question-answering interfaces.

The central problem is not only that LLMs hallucinate. In this paper, a *source-support hallucination* occurs when a model cites an existing authority but the cited text does not support the proposition asserted. That category includes converting guidance into obligation, generalizing a regional rule as universal law, erasing treaty exceptions, or inflating soft law into peremptory law. Empirical studies show fabricated or misleading authority even in RAG-supported legal products [37–40]. LegalBench and LegalBench-RAG provide important methodological support for decomposing legal reasoning and evaluating legal retrieval, but they leave open the domain-specific checks needed here: source hierarchy, deontic modality, actor and jurisdictional scope, and recursive memory contamination [35, 36, 50, 51]. LegalBench measures broad legal-reasoning tasks, LegalBench-RAG measures retrieval of legally relevant passages, evidence-centered benchmark design specifies construct-validity discipline, and normative-reasoning benchmarks isolate modal reasoning; none by itself tests whether a modern slavery claim preserves all of source authority, addressee, jurisdiction, exception entailment, and memory eligibility. Those open checks are not merely technical metrics. They reflect international-law doctrine, where treaties, reports, interpretations, regional instruments, soft law, and scholarship do not form a flat hierarchy [31–34].

This paper argues that social LLM systems used for modern slavery reasoning need a verifier-governed architecture that treats legal propositions as structured claims before they become prose and before they enter memory. The proposed architecture, Recursive Norm Refinement (RNR), has four commitments. First, it decomposes outputs into claim-level units with fields for source, actor, legal regime, jurisdiction, temporal scope, modality, exception, and evidence. Second, it verifies each field against a curated hierarchy of primary sources and authoritative interpretations. Third, it uses disagreement and abstention as safety mechanisms when source support is incomplete, contested, or outside scope. Fourth, it gates recursive memory so that unverified claims are quarantined rather than

reused as premises in later cycles. Human-rights impact assessment and algorithmic auditing supply the governance workflow for affected-rights mapping, severity, mitigation, documentation, and accountability [18–21, 25, 24]. Legal RAG, process supervision, factuality evaluation, and self-refinement research supply the engineering pattern for retrieval, claim decomposition, step-level verification, and revision [46, 47, 53, 52].

The paper formalizes modern slavery reasoning as a source-sensitive task, specifies the RNR pipeline, and reports three reproducible diagnostic experiments. The experiments use a deterministic adversarial claim generator rather than external model calls, isolating verifier behaviour from a proprietary model snapshot.

2 Modern Slavery Law as a Source-Sensitive Reasoning Task

A social LLM system reasoning about modern slavery must not merely detect words such as *trafficking*, *forced labour*, or *slavery*. It must preserve legal distinctions that matter for prevention, remedy, liability, and institutional design. Table 1 summarizes the core requirements that the architecture operationalizes.

The first source-sensitive distinction is between forced labour and slavery. Forced labour under ILO Convention No. 29 is broader than slavery because it does not require exercise of powers attaching to ownership [2, 4]. A system that converts every forced-labour signal into "slavery" may inflate a legal claim; a system that treats slavery only as historic chattel slavery may miss contemporary ownership-like control. The Supplementary Convention expands attention to institutions and practices similar to slavery, including debt bondage, serfdom, servile marriage, and exploitation of children or young persons [5]. The Palermo Protocol then connects trafficking to acts, means, and exploitative purposes; trafficking may be present even when the final exploitative condition is not yet fully realized [7].

The second distinction is between universal, regional, and sector-specific instruments. Article 8 ICCPR prohibits slavery, servitude, and forced or compulsory labour, with enumerated exclusions [6]. Article 4 ECHR has a similar structure within the Council of Europe system [8]. Article 5 of the EU Charter expressly prohibits slavery, servitude, forced labour, and trafficking, but it binds EU institutions and Member States when they implement EU law [9]. Article 5 of the African Charter protects dignity and prohibits exploitation and degradation, including slavery and slave trade [10]. Article 10 of the Arab Charter prohibits slavery, servitude, forced labour, and trafficking [11]. These instruments are mutually reinforcing, but they are not interchangeable.

The third distinction is between individual criminal liability, state responsibility, and business responsibility. The Rome Statute treats enslavement and deportation or forcible transfer as crimes against humanity when contextual elements are met [13]. Geneva Convention IV prohibits forcible transfers and deportations of protected persons from occupied territory [12]. The UN Guiding

Table 1. Legal requirements translated into verifier tests.

Requirement	Legal significance	Computational test
Source hierarchy	A treaty obligation, official soft-law expectation, court interpretation, and scholarly claim have different legal force [31–34].	Classify authority type, binding status, issuing institution, and relationship to primary sources.
Deontic modality	"Must", "must not", "should", "may", and definitional provisions do different legal work [51].	Extract and compare modality against the cited provision; reject unsupported upgrades or downgrades.
Actor and addressee	States, occupying powers, businesses, deployers, providers, individuals, and international organizations are not interchangeable; the relevant instrument must name or otherwise reach the actor [12–14, 16, 17].	Check whether the cited norm addresses the actor named in the claim.
Jurisdiction and scope	EU, Council of Europe, African, Arab, ILO, ICC, and universal instruments have distinct territorial, personal, and temporal scopes [2, 6, 8, 10, 11, 13, 16, 17].	Store jurisdiction, temporal applicability, and treaty-party or implementation conditions.
Exception entailment	Several forced-labour prohibitions contain exclusions for military service, civic obligations, detention work, or emergencies [2, 6, 8].	Reject universal claims unless the cited source entails that no exception or limitation applies.
Remedy and due diligence	Business responsibilities and AI impact assessments are often process obligations rather than strict outcome guarantees [14–16].	Distinguish prevention, mitigation, remediation, documentation, and sanction duties.

Principles on Business and Human Rights (UNGPs) articulate the corporate responsibility to respect human rights through due diligence and remediation; they are authoritative soft law, not a treaty imposing identical obligations on every firm [14]. In technology settings, OHCHR's B-Tech work translates the UNGPs to technology companies, while the EU AI Act and the Council of Europe AI Framework Convention create or anticipate more formalized impact-assessment obligations for specified actors and systems [15–17].

The fourth distinction is temporal and evidentiary. International law changes through treaty ratification, amendment, institutional interpretation, and practice. Soft law can influence hard law and vice versa, but it cannot be cited as though it were automatically binding treaty law [31, 32]. Human rights interpretation can evolve, but evolutive interpretation does not license a model to invent ungrounded duties. In practice, a modern slavery risk answer should therefore store not just a conclusion, but also why the conclusion follows from the cited source and what would defeat or narrow it.

3 Architecture: Recursive Norm Refinement

RNR is an agent architecture for high-stakes legal reasoning in social computing systems. It combines retrieval, claim decomposition, external verification, critique, revision, and memory gating. The model can still generate natural language, but its legally consequential propositions are treated as structured objects. Figure 1 shows the pipeline.

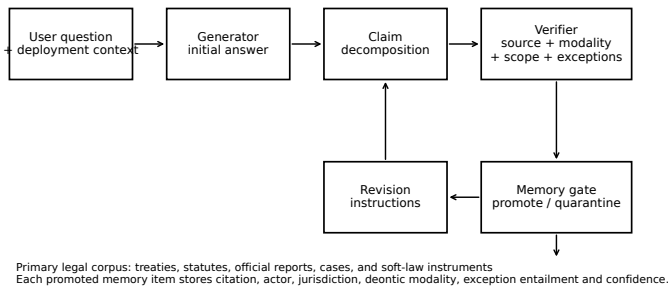

Fig. 1. Verifier-governed recursive norm refinement. The language model drafts and revises, but promotion to memory is conditioned on source, modality, scope, and evidence checks.

The legal corpus is organized by authority rather than by semantic similarity alone. Each source record stores title, provision, instrument type, authority tier, legal regime, binding status, addressee, jurisdictional scope, temporal/version

field, predicate, exception text, and keywords. Ranking first filters by actor, jurisdiction, and time, then applies a fixed tier order: peremptory or non-derogable core when independently established; binding treaties, regulations, protocols, and charters; authoritative institutional interpretations and official reports; official soft law; peer-reviewed scholarship; and encyclopedic sources as finding aids. Primary instruments therefore outrank summaries; scholarly and encyclopedic materials can guide interpretation, but the verifier tries to cite the original treaty, case, statute, report, or institutional instrument whenever possible.

The claim schema is the atomic unit. A claim is not simply a sentence; it is a tuple:

$$\mathsf{claim} = (p, s, a, j, t, d, v, x, c),$$

where p is the proposition, s the cited source, a the actor, j the jurisdiction or legal regime, t the temporal scope, d the deontic modality, v the evidence span, x the exception or limitation, and c a confidence state. The x field is an entailment check, not a label: a claim that asserts universal or exception-free application is accepted only if the cited source entails that form; otherwise it is narrowed, rejected, or abstained. This structure follows factuality evaluation, legal benchmarking, and process supervision, where final answers are only as reliable as their weakest unsupported claim [46–48, 53]. It also reflects the legal literature's insistence that source hierarchy, indeterminacy, and pluralism are not optional metadata [33, 34].

Algorithm 1: claim verification	**Algorithm 2: memory promotion**
1. Decompose draft answer into candidate claims.	1. Receive verified, rejected, and ambiguous claims.
2. Extract source, actor, jurisdiction, time, modality, exception, and evidence span.	2. Require citation, source type, actor, scope, modality, date, and evidence span before promotion.
3. Retrieve candidate authorities using semantic, lexical, structural, and citation anchors.	3. If a claim conflicts with memory, open a conflict record rather than overwriting.
4. Rank authorities by source hierarchy; prefer original primary sources over summaries.	4. Promote only if confidence exceeds threshold and no unresolved higher-authority contradiction remains.
5. Test actor, scope, modality, and exception entailment.	5. Store rejected or ambiguous claims in quarantine; retrieve them later only as warnings.
6. Return verified, narrowed, unsupported, or abstained claims with an audit record.	6. Use promoted claims as premises in later cycles.

Fig. 2. Verifier-gated refinement and memory-control procedure. Unsupported and ambiguous claims do not become premises for later reasoning.

Figure 2 differs from ordinary RAG. A standard RAG pipeline can retrieve relevant passages yet still produce a legally invalid answer by overstating modality or scope. The verifier therefore checks claim fields, not only passage similarity. This is consistent with work showing that legal reasoning tasks require more than general language understanding [42, 43, 41]. It also aligns with structure-

aware legal RAG, which represents legal norms by hierarchy, temporality, and provenance rather than arbitrary chunks [44, 45].

Recursive memory is also essential. Self-refinement and Reflexion-style systems demonstrate that language agents can improve by storing feedback and using it in later cycles [52, 54]. But legal memory has a failure mode: an unsupported legal claim can harden into an apparently established premise. Huang and colleagues show that ungrounded self-correction can worsen reasoning without external feedback [55]. In legal systems, the external feedback must include source verification.

This architecture does not assume that all legal uncertainty can be eliminated. International law often permits plausible competing interpretations, conflicts among specialized regimes, and soft-law instruments with real normative significance but no formal treaty force [31–34]. RNR therefore uses abstention, narrowing, and conflict records as positive outputs. For example, a claim that technology companies "must" remediate all modern slavery harms everywhere is too broad under the UNGPs; a narrower claim that companies should provide for or cooperate in remediation where they caused or contributed to adverse impacts is source-faithful [14]. Similarly, a claim that every high-risk AI system in the world requires a legally binding FRIA is too broad; Article 27 of the EU AI Act applies to specified deployers and specified high-risk systems in the EU framework [16, 21].

4 Benchmark Design and Diagnostic Experiments

The experiments test the verifier and memory gates under controlled adversarial conditions; they do not claim that a particular commercial or open-source LLM has the measured scores. Decoupling the verifier evaluation from a specific model snapshot ensures reproducibility as model APIs, system prompts, retrieval indices, and safety policies change. The experiment package released with this paper generates all claims, scenarios, metrics, and figures from local code.

The benchmark design follows three principles from evidence-centered benchmark design and factuality evaluation [50, 46]. First, the construct is explicitly defined: source-faithful modern slavery legal reasoning requires support, modality, actor, scope, exception entailment, and authority classification. Second, tasks are designed to elicit failures in that construct: modality swaps, source inflation, scope drift, actor drift, wrong-source citation, absolute overstatement, and over-generalization. Third, metrics distinguish false acceptance, supported-claim recall, and abstention on ambiguous claims rather than collapsing all performance into one holistic score.

A small natural-prose smoke test was also used to check that the verifier handles paragraph-style outputs, not only templated claims. Three short draft answers were written to resemble ordinary LLM risk summaries: one overstated the UNGPs as binding treaty law, one generalized the EU AI Act FRIA duty to all global AI deployers, and one paraphrased ILO Convention No. 29 while preserving the menace-of-penalty and voluntariness elements. The verifier rejected

the first two after actor, authority, and jurisdiction checks and accepted the third only after preserving the Convention's definitional scope. This qualitative check is not included in the quantitative metrics; it functions as a sanity test for applying the claim schema to prose.

Table 2. Adversarial legal-error taxonomy used in Experiment 1.

Error type	Example pattern	Why it matters
Modality swap	A source saying "should" is reported as "must", or a prohibition is converted into permission.	Changes the legal force of the claim.
Source inflation	A soft-law or ordinary treaty claim is described as peremptory, absolute, or binding without support.	Misleads downstream compliance and remedy decisions.
Scope drift	A regional or party-specific rule is generalized to all states, firms, or AI systems.	Erases jurisdiction and treaty-party limits.
Actor drift	A duty of states is attributed to companies, providers, deployers, or individuals, or vice versa.	Misidentifies the responsible institution.
Wrong-source citation	A real source is cited for a proposition grounded elsewhere.	Looks credible while being unsupported.
Absolute overstatement	A claim preserves the source but erases exceptions, limitations, or contextual thresholds.	Produces a plausible but legally unsafe rule.
Over-generalization	A broad risk slogan is asserted without sufficient legal anchoring.	Should trigger narrowing or abstention.

4.1 Experiment 1: Source and Modality Verification

Experiment 1 uses 25 source clauses from original legal instruments and official AI/human-rights materials. The generator creates 181 candidate claims: supported claims, supported paraphrases, unsupported perturbations, and ambiguous over-generalizations. We compare four policies: retrieval-only lexical overlap, source-only existence, source+deontic modality, and full RNR citation, actor, scope, modality, exception-entailment, and ambiguity checks.

Figure 3 reports headline metrics. Full RNR achieved 1.000 overall accuracy in this controlled set, rejected all unsupported claims, recalled all supported claims, and abstained on all ambiguous over-generalizations. The source+deontic

ablation improved over source-only verification but still falsely accepted many actor, scope, and absolute-overstatement errors. Source-only verification performed poorly because the adversarial claims cite real instruments while changing the legal proposition.

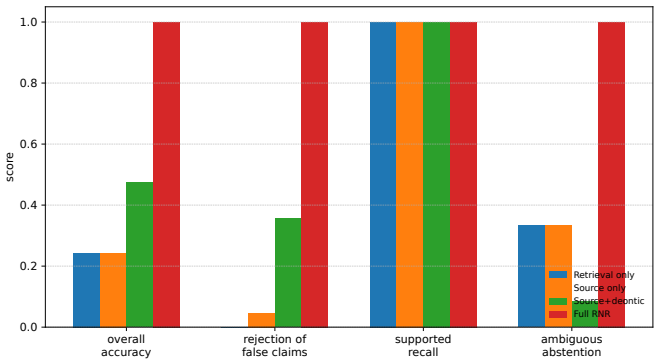

Fig. 3. Experiment 1 metrics across verifier ablations. Rejection of false claims is one minus the false-accept rate.

The error-type breakdown in Figure 4 explains why the full schema is necessary: deontic checking catches modality swaps, but it cannot catch scope drift or actor drift. Source-only checking is especially vulnerable to source inflation: a sentence can cite the right treaty and still say the wrong thing about the treaty's force. The added exception-entailment rule rejects absolute-overstatement claims that preserve shallow fields while adding a "no exception or limitation" proposition not entailed by the source.

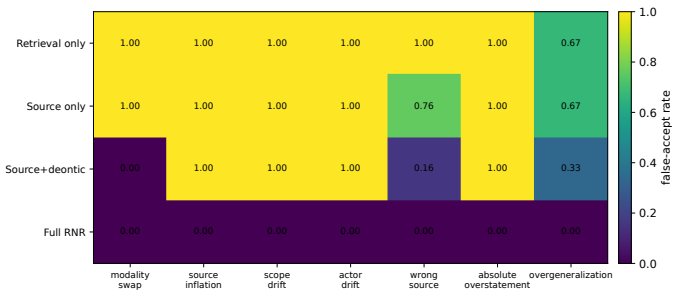

Fig. 4. False-accept rate by error type. Full RNR catches perturbations in this controlled set after explicit exception-entailment checking for absolute overstatements.

4.2 Experiment 2: Recursive Memory Contamination

Experiment 2 simulates 30 recursive cycles. In each cycle, candidate claims are sampled and may be promoted to memory under three policies: no gate, source-only gate, and full RNR gate. Later cycles sample partly from prior memory, modelling how agents reuse stored premises. The outcome is the share of promoted memory claims that are unsupported or ambiguous.

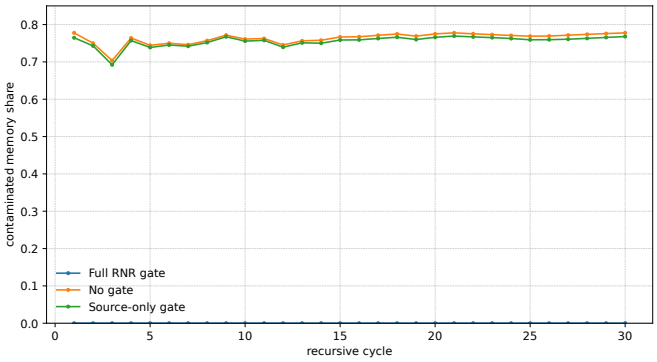

Fig. 5. Experiment 2 memory contamination over recursive cycles. Source-only gating barely reduces contamination because adversarial claims are constructed to cite real sources while changing the legal proposition.

Figure 5 illustrates the main safety claim. Without a gate, contamination rapidly dominates memory. A source-only gate performs only slightly better because many false claims cite real sources. With exception entailment enabled, full RNR reduces final contamination to 0.000 in this controlled run. The result should not be read as field accuracy; it shows that memory safety follows only when source existence is combined with actor, scope, modality, and exception checks.

4.3 Experiment 3: Rights-Aware Triage for AI-Mediated Labour-Risk Scenarios

Experiment 3 examines whether a rights-aware scoring rule changes prioritization of modern slavery risks in social computing use cases. We generated 48 stylized scenarios across recruitment, supply-chain audit, platform work, migration screening, content moderation, and welfare targeting. Each scenario receives features for coercion signal, vulnerability, scale, reversibility, and evidence quality. A flat risk score averages scale, reversibility, and evidence quality. A rights-aware score gives explicit weight to coercion and vulnerability. These numerical weights are experimental design choices, not weights prescribed by the

cited frameworks; they are inspired by HRIA, FRIA, and algorithmic-impact-assessment scholarship on severity, affected groups, vulnerability, reversibility, mitigation, documentation, and remedy [20–23, 25].

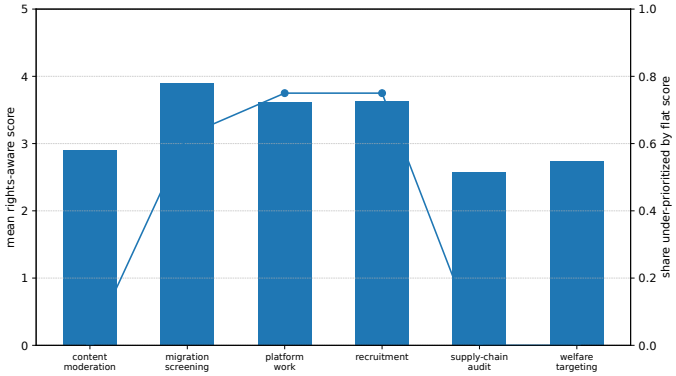

Fig. 6. Experiment 3 scenario triage. A flat operational score under-prioritizes several coercion-heavy domains; the rights-aware score elevates recruitment, migration, and platform-work scenarios.

The flat score under-prioritized 0.708 of scenarios labelled severe by the rights-aware rule. The gap is largest in recruitment, platform work, and migration screening, where coercion and vulnerability can be high even when evidence quality is incomplete. This is consistent with the cited impact-assessment literature as an analytic design principle rather than a prescribed formula: severe rights interferences should not be hidden by sparse evidence or ordinary operational risk averages [19, 22]. For RNR, the lesson is that the verifier should not only ask whether a claim is source-supported; it should also require transparent documentation of why a modern slavery risk was triaged as severe, moderate, or unsupported.

5 Discussion

The experiments support a practical claim: legal risk reasoning in social LLM systems should be evaluated at the claim level, not only at the answer level. A final paragraph may read well while one sentence misstates the law. In modern slavery contexts, a false statement that a corporate guideline is binding treaty law, that a forced-labour prohibition admits no exception, or that an EU deployer duty applies globally can distort accountability. Underclaiming is also dangerous: if a model treats trafficking as only completed exploitation, it may miss recruitment-stage coercion.

RNR is also a response to limitations in AI governance itself. AI ethics guidelines show convergence around principles but divergence in implementation [26].

Critics argue that principles alone do not guarantee accountability and may invite ethics washing [27, 28]. Risk-based regulation adds structure, but scholarship on the EU AI Act cautions that risk categories, conformity assessment, redress, and fundamental-rights protection require careful design [29, 30]. A social LLM system that merely retrieves documents and drafts plausible prose would reproduce this gap. In the terms of ethics-washing critiques, it may perform surface compliance while failing to preserve the legal predicates that make accountability possible.

HRIA and FRIA methodologies ask organizations to identify affected rights, stakeholders, severity, mitigation, documentation, and remedy [19–21]. RNR provides a machine-readable substrate for those steps by converting assertions into auditable claims; the legal and participatory judgment required by HRIA remains outside the verifier itself. For example, a supply-chain monitoring system could store a claim that a recruitment-screening model creates elevated forced-labour risk because it interacts with migrant-worker debt vulnerability, cites the relevant legal basis, identifies the affected group, records uncertainty, and recommends human review. If later evidence changes, the claim can be revised without erasing the audit trail.

The architecture has limits. The experiments are constructed rather than field deployments, and legal entailment is harder than lexical comparison. The exception-entailment rule handles explicit exception erasure in this controlled set, but implied limitations, proportionality standards, and contested derogations still require expert validation. Source hierarchy also does not eliminate interpretive disagreement, and the architecture assumes a maintained legal corpus with accurate versioning metadata and authority-tier review.

These limits suggest a research agenda. Legal RAG benchmarks should move beyond document retrieval to claim-level source fidelity, deontic modality, and scope preservation; this is an extension of retrieval and citation evaluation, not a claim that existing legal RAG benchmarks already measure all those constructs [47, 36, 51]. Multilingual benchmarks are essential because modern slavery risk is global [49]. Verifier-governed agents should surface contested interpretations before memory promotion because international-law argument often depends on source authority, regime conflict, and institutional competence [31–34]. Recursive-agent and process-supervision methods provide the engineering pattern for doing so, while evidence-centered benchmark design supplies the construct-validity discipline [50, 52, 53].

6 Conclusion

Modern slavery risk reasoning is a demanding test case for responsible social computing. It involves vulnerable populations, multi-institutional data flows, transnational legal sources, and severe downstream consequences. LLM agents can help by organizing evidence and drafting analyses, but only if their legal claims are grounded, scoped, and auditable. Recursive Norm Refinement provides one operational method for doing so. It decomposes answers into struc-

tured claims, verifies each claim against original legal sources and source hierarchy, checks modality, scope, and exception entailment, surfaces ambiguity, and gates memory promotion. The diagnostic experiments show that retrieval and citation existence are not enough; legally meaningful verification must inspect the fields that make a proposition valid. The approach favours conservatism: in modern slavery contexts, a system that narrows or abstains when source support is uncertain is not less useful than a fluent system that overstates the law. It is safer, more accountable, and better aligned with the legal and social stakes of the domain.

Disclosure of Interests. The author(s) have no competing interests to declare that are relevant to the content of this article.

References

1. International Labour Organization, Walk Free, International Organization for Migration: Global Estimates of Modern Slavery: Forced Labour and Forced Marriage. ILO, Geneva (2022).
2. International Labour Organization: Forced Labour Convention, 1930 (No. 29). 39 UNTS 55 (1930).
3. International Labour Organization: Abolition of Forced Labour Convention, 1957 (No. 105). 320 UNTS 291 (1957).
4. League of Nations: Slavery Convention. 60 LNTS 253 (1926).
5. United Nations: Supplementary Convention on the Abolition of Slavery, the Slave Trade, and Institutions and Practices Similar to Slavery. 266 UNTS 3 (1956).
6. United Nations General Assembly: International Covenant on Civil and Political Rights. 999 UNTS 171 (1966).
7. United Nations General Assembly: Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children. 2237 UNTS 319 (2000).
8. Council of Europe: Convention for the Protection of Human Rights and Fundamental Freedoms. ETS No. 5 (1950).
9. European Union: Charter of Fundamental Rights of the European Union. 2012 OJ C 326/391 (2000/2012).
10. Organization of African Unity: African Charter on Human and Peoples' Rights. 1520 UNTS 217 (1981).
11. League of Arab States: Arab Charter on Human Rights. Reprinted in 12 Int. Hum. Rts. Rep. 893 (2004).
12. Diplomatic Conference of Geneva: Geneva Convention Relative to the Protection of Civilian Persons in Time of War. 75 UNTS 287 (1949).
13. United Nations Diplomatic Conference: Rome Statute of the International Criminal Court. 2187 UNTS 90 (1998).
14. United Nations Human Rights Council: Guiding Principles on Business and Human Rights: Implementing the United Nations Protect, Respect and Remedy Framework. UN Doc. A/HRC/17/31 (2011).
15. OHCHR B-Tech Project: The UN Guiding Principles on Business and Human Rights: An Introduction for Technology Companies. OHCHR (2020).
16. European Parliament and Council: Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence. Official Journal of the European Union (2024).

17. Council of Europe: Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. CETS No. 225 (2024).
18. McGregor, L., Murray, D., Ng, V.: International Human Rights Law as a Framework for Algorithmic Accountability. Int. Comp. Law Q. 68(2), 309–343 (2019).
19. Mantelero, A.: AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment. Comput. Law Secur. Rev. 34(4), 754–772 (2018).
20. Mantelero, A., Esposito, M.S.: An Evidence-Based Methodology for Human Rights Impact Assessment in the Development of AI Data-Intensive Systems. Comput. Law Secur. Rev. 41, 105561 (2022).
21. Mantelero, A.: The Fundamental Rights Impact Assessment in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template. Comput. Law Secur. Rev. 54, 106020 (2024).
22. Malgieri, G., Santos, C.: Assessing the Severity of Impacts on Fundamental Rights. Comput. Law Secur. Rev. 56, 106113 (2025).
23. Stahl, B.C., Antoniou, J., Ryan, M., et al.: A Systematic Review of Artificial Intelligence Impact Assessments. Artif. Intell. Rev. 56, 12799–12831 (2023).
24. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., et al.: Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In: FAT* 2020, pp. 33–44 (2020).
25. Metcalf, J., Moss, E., Watkins, E.A., Singh, R., Elish, M.C.: Governing Algorithmic Systems with Impact Assessments: Six Observations. In: AIES 2021, pp. 191–200 (2021).
26. Jobin, A., Ienca, M., Vayena, E.: The Global Landscape of AI Ethics Guidelines. Nat. Mach. Intell. 1, 389–399 (2019).
27. Mittelstadt, B.: Principles Alone Cannot Guarantee Ethical AI. Nat. Mach. Intell. 1, 501–507 (2019).
28. Bietti, E.: AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics. Digit. Soc. 1, 9 (2022).
29. Veale, M., Zuiderveen Borgesius, F.: Demystifying the Draft EU Artificial Intelligence Act. Comput. Law Rev. Int. 22(4), 97–112 (2021).
30. Smuha, N.A.: Normative Challenges of Risk Regulation of Artificial Intelligence. Ethics Inf. Technol. 26, 28 (2024).
31. Boyle, A.: Some Reflections on the Relationship of Treaties and Soft Law. Int. Comp. Law Q. 48(4), 901–913 (1999).
32. Shaffer, G.C., Pollack, M.A.: Hard vs. Soft Law: Alternatives, Complements, and Antagonists in International Governance. Minn. Law Rev. 94, 706–799 (2010).
33. International Law Commission Study Group: Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law. A/CN.4/L.682 (2006).
34. Koskenniemi, M.: From Apology to Utopia: The Structure of International Legal Argument. Cambridge University Press, Cambridge (2005).
35. Guha, N., Nyarko, J., Ho, D.E., Re, C., et al.: LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. NeurIPS Datasets and Benchmarks (2023).
36. Pipitone, N., Houir Alami, G.: LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. arXiv:2408.10343 (2024).
37. Dahl, M., Magesh, V., Suzgun, M., Ho, D.E.: Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. J. Legal Anal. 16(1), laae003 (2024).
38. Magesh, V., et al.: Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. arXiv:2405.20362 (2024).

39. Hou, A.B., et al.: Gaps or Hallucinations? Gazing into Machine-Generated Legal Analysis for Fine-Grained Text Evaluations. arXiv:2409.09947 (2024).
40. Chen, S., Yin, H., Zhou, S.: LegalCiteBench: Evaluating Citation Reliability in Legal Language Models. arXiv:2605.10186 (2026).
41. Linna, E., Linna, T.: Challenges for Generative AI in Legal Reasoning. Discover Artificial Intelligence 6, 170 (2026).
42. Holzenberger, N., Blair-Stanek, A., Van Durme, B.: A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering. In: ACL 2020, pp. 5658–5668 (2020).
43. Blair-Stanek, A., Holzenberger, N., Van Durme, B.: Can GPT-3 Perform Statutory Reasoning? In: ICAIL 2023, pp. 22–31 (2023).
44. de Martim, H.: An Ontology-Driven Graph RAG for Legal Norms: A Structural, Temporal, and Deterministic Approach. Frontiers in Artificial Intelligence and Applications (2025).
45. de Martim, H.: Deterministic Legal Agents: A Canonical Primitive API for Auditable Reasoning over Temporal Knowledge Graphs. arXiv:2510.06002 (2025).
46. Min, S., Krishna, K., Lyu, X., et al.: FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In: EMNLP 2023, pp. 12076–12100 (2023).
47. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling Large Language Models to Generate Text with Citations. In: EMNLP 2023, pp. 6465–6488 (2023).
48. Es, S., James, J., Espinosa-Anke, L., Schockaert, S.: RAGAS: Automated Evaluation of Retrieval Augmented Generation. EACL 2024 (2023).
49. Niklaus, J., Matoshi, V., Rani, P., et al.: LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. In: Findings of EMNLP 2023, pp. 3016–3052 (2023).
50. Liu, Y.L., Blodgett, S.L., Cheung, J.C.K., Liao, Q.V., Olteanu, A., Xiao, Z.: ECBD: Evidence-Centered Benchmark Design for NLP. In: ACL 2024 (2024).
51. Ozeki, K., Ando, R., Morishita, T., Abe, H., Mineshima, K., Okada, M.: Normative Reasoning in Large Language Models: A Comparative Benchmark from Logical and Modal Perspectives. In: BlackboxNLP 2025 (2025).
52. Madaan, A., Tandon, N., Gupta, P., et al.: Self-Refine: Iterative Refinement with Self-Feedback. NeurIPS 2023 (2023).
53. Lightman, H., Kosaraju, V., Burda, Y., et al.: Let's Verify Step by Step. ICLR 2024 (2023).
54. Shinn, N., Cassano, F., Berman, E., et al.: Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS 2023 (2023).
55. Huang, J., Chen, X., Mishra, S., et al.: Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024 (2023).