Copyright Infringement by Large Language Models in the EU: Misalignment, Safeguards, and the Path Forward

Noah Scharrenberg
Contractuo
Maastricht University
noah@contractuo.com

Chang Sun
Maastricht University
chang.sun@maastrichtuniversity.nl

Abstract

This position paper argues that European copyright law has struggled to keep pace with the development of large language models (LLMs), possibly creating a fundamental epistemic misalignment: copyright compliance relies on qualitative, context-dependent standards, while LLM development is governed by quantitative, proactive metrics. This gap means that technical safeguards, by themselves, may be insufficient to reliably demonstrate legal compliance. We identify several practical limitations in the existing EU legal frameworks, including ambiguous "lawful access" rules, fragmented opt-outs, and vague disclosure duties. We then discuss technical measures such as provenance-first data governance, machine unlearning for post-hoc removal, and synthetic data generation, showing their promise but also their limits. Finally, we propose a path forward grounded in legal-technical co-design, suggesting directions for standardising machine-readable opt-outs, disclosure templates, clarifying core legal terms, and developing legally-informed benchmarks and evidence standards. We conclude that such an integrated framework is essential to make compliance auditable, thus protecting creators' rights while enabling responsible AI innovation at scale.

1 Introduction

In 2025, researchers demonstrated that Meta's LLaMA 3.1 could reproduce entire chapters of copyright-protected novels, including *Harry Potter*, almost verbatim (Cooper et al., 2025). In Germany, the case of *Kneschke v. LAION* exposed the fragility of text-and-data mining (TDM) safeguards when a photographer's images were scraped into a large-scale dataset without consent (Hamburg District Court, 2024; Havlíková, 2025). These incidents are indicative of broader foreseeable risks emerging from how large language models (LLMs) are trained: by ingesting petabytes of text, images, and code from the open Internet, much of it copyright-protected (Borhi et al., 2025; Quintais, 2025; Margoni and Kretschmer, 2022).

The European Union (EU) has responded with a layered legal framework, most prominently Copyright in the Digital Single Market (CDSM) (European Parliament and Council, 2019) Directive and the AI Act (European Parliament and Council, 2024). Rooted in the civil law tradition, this framework relies on defined exceptions and limitations, contrasting with the broader and more flexible "fair use" doctrine in common law jurisdictions such as the United States. The instruments establish the principles of lawful access, opt-out mechanisms, and transparency obligations (European Parliament and Council, 2019, 2024). However, they were not designed to accommodate the technical realities of large-scale model development, including petabyte-scale data ingestion or probabilistic memorisation (Borhi et al., 2025; Quintais, 2025). Whereas copyright law is structured to resolve human-scale disputes retrospectively through courts; LLMs operate at machine-scale, enabling the ingestion and reproduction of billions of works, with minimal prompting (Borhi et al., 2025). The mismatch contributes to a gap between legal expectations and technical characteristics of generative AI (e.g., LLMs) systems.

This practical mismatch reflects a deeper structural problem. Copyright law relies on qualitative, context-dependent standards (originality, substantial similarity) adjudicated retrospectively by courts (European Parliament and Council, 2019; Court of Justice of the European Union, 2009, 2019), whilst LLM development operates through quantitative, proactive metrics (loss functions, similarity thresholds) optimised by automated pipelines (Chen et al., 2024; Wei et al., 2024). These domains operate in different conceptual languages, making it impractical to demonstrate legal compliance through technical metrics alone. This creates a persistent risk of partial or structural

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 125–134
November 8, 2025 ©2025 Association for Computational Linguistics

non-compliance: developers cannot credibly prove copyright respect, rightsholders cannot meaningfully enforce their rights, and regulators lack enforceable technical standards (Borhi et al., 2025; Buick, 2024).

We suggest that effective copyright compliance in the context of LLMs requires interdisciplinary co-design, where law frameworks inform the development of technical metrics and technical methods support compliance verification at scale. To support this position, this paper looks into why current EU frameworks fall short when applied to LLMs, highlighting the mismatch between retrospective legal adjudication and proactive technical safeguards. We then analyse the structural misalignment between qualitative legal standards and quantitative metrics before surveying the promise and limits of current technical solutions. Finally, we propose a path forward: a triad of provenance infrastructure, adversarially robust unlearning, and clean-chain synthetic data generation, all embedded within a framework of legally-informed evaluation.

2 Current Legal and Technical Frameworks for Generative AI

The EU's legal response to generative AI is layered. The CDSM Directive created TDM exceptions with opt-out rights for rightsholders, and the AI Act added transparency and risk-management obligations for general-purpose AI (GPAI) (European Parliament and Council, 2019, 2024). Taken separately, each instrument is internally coherent. Taken together, they reveal practical limitations for large-scale model training and deployment.

2.1 Legal Gaps That Matter in Practice

EU copyright law harmonises certain economic rights (notably reproduction and communication to the public) whilst leaving others, including moral rights and specific limitations, partially within Member State competence (European Parliament and Council, 2001, 2019). Similarly, whilst Article 3 and Article 4 CDSM establish mandatory TDM exceptions, Member States retain discretion in implementing supplementary provisions and procedural mechanisms (European Parliament and Council, 2019; Margoni and Kretschmer, 2022). This partial harmonisation creates compliance complexity for cross-border AI development, as developers have to navigate both EU-level directives and nationally divergent implementations.

Lawful access is ambiguous. Article 4 of the CDSM Directive permits TDM on "lawfully accessible" works unless rights are expressly reserved (European Parliament and Council, 2019). However, "lawful access" is left ambiguous for developers: *Does it require compliance with the website terms of service (ToS)? Can contractual prohibitions in a licence defeat this statutory exception? Does accessing content behind a paywall suffice?* As recent litigations and analysis show, the lack of a clear definition can turn compliance into high-stakes guesswork for developers and leaves rightsholders without a stable enforcement baseline (Quintais, 2025; Hamburg District Court, 2024; Dermawan, 2024).

As suggested by (Margoni and Kretschmer, 2022) and (Quintais, 2025), this ambiguity appears to be intentional, designed to preserve Member State flexibility in implementation, but this flexibility becomes a liability when governing automated systems that require clear, consistent signals. Some commentators argue for a more expansive interpretation that would facilitate AI development, whilst others advocate for strict construction to protect rightsholder interests (Dermawan, 2024; Havlíková, 2025).

Reservations "in an appropriate manner". Article 4(3) CDSM allows rightsholders to reserve their reproduction and extraction rights against TDM under Article 4(1)-(2), thus disabling the statutory TDM exception for their works. The Directive, however, does not standardise machine-readable signalling or define what counts as an *appropriate* reservation across contexts (web, platforms, feeds, datasets). Commentary and practice show that *robots.txt*, HTTP headers, meta tags, or natural-language ToS coexist and are often brittle for automated ingestion (Margoni and Kretschmer, 2022; Hamann, 2024; Keller, 2024). Absent a harmonised schema or registry, developers face non-exhaustive signals that do not scale reliably to petabyte-level pipelines (European Commission, DG CNECT, 2025).

Vague transparency duties. The AI Act requires GPAI providers to publish a "sufficiently detailed summary" of training content (European Parliament and Council, 2024). However, what qualifies as "sufficiently detailed" remains open. High-level labels like "web crawl" are of limited use to rightsholders because they lack work- or domain-level traceability needed for verifying reservations and targeted takedowns; granular disclosure raises trade-secret concerns and is technically burdensome. Without clear templates and thresholds, transparency risks becoming largely symbolic rather than meaningfully verifiable (Quintais, 2025; Buick, 2024; Warso and Gahntz, 2024).

Jurisdictional and temporal gaps. Training can be geographically and temporally distributed among vendors, regions, and versions. The CDSM Directive/AI Act connection focusses on models placed on the EU market, but provides limited tools to assess where and how reproduction occurred or to retrospectively correct legacy training (European Parliament and Council, 2019, 2024; Quintais, 2025; Lucchi and Hunter, 2025).

2.2 Technical Gaps That Surface at the Output Stage Under Existing Law

Output-side memorisation and beyond. Verbatim regurgitation is no longer rare. It correlates with data repetition, model capacity, and weak safeguards (Cooper et al., 2025; Chen et al., 2024). More subtle risks such as substantial similarity, plot and character appropriation, or unauthorised derivative works—are harder to detect and measure with current toolchains (Chen et al., 2024; Russinovich and Salem, 2025; Chun, 2024).

Opaqueness and non-determinism Model internals and training recipes are opaque, and outputs are probabilistic (Quintais, 2025; Borhi et al., 2025). Even well-intentioned providers often cannot reliably prove a negative (that a given work was not in training, and assuming that a teacher model or an existing pre-trained model was used) or guarantee the absence of infringing outputs under adversarial prompting (Wei et al., 2024; Jin et al., 2024; Shi et al., 2024).

Metric—standard mismatch. Existing technical controls optimise quantitative metrics (e.g., ROUGE, cosine similarity, LCS) that do not consistently map to qualitative legal tests (e.g., substantial similarity as a holistic impression, market substitution effects). Optimising the former does not by itself ensure compliance with the latter (Cooper et al., 2025; Chen et al., 2024; Wei et al., 2024; Chun, 2024).

2.3 Why Internally Coherent Systems Still Misfire Together

Both legal and technical regimes make sense on their own terms. Copyright law is built for human-scale, retrospective adjudication: a work, defendant, a court, and a remedy (Lucchi and Hunter, 2025; Quintais, 2025). LLM development is built for machine-scale, proactive control: billions of files, automated ingestion, and statistical learning that must be governed prospectively (Borhi et al., 2025; European Commission, DG CNECT, 2025).

This fundamental disconnect creates what we might term a "compliance impossibility tension". Doctrines built for retrospective, human-scale adjudication are ill-suited to govern automated systems that require proactive, machine-scale controls. Developers are left without clear, machine-actionable constraints, while rightsholders cannot reliably audit compliance at scale. This could require a shift towards co-designed standards that are legally meaningful and technically implementable (Borhi et al., 2025; Quintais, 2025; Lucchi and Hunter, 2025), an issue rooted in the epistemic misalignment explored in the section 3.

3 Qualitative Law vs. Quantitative Metrics

At the core of the compliance challenge lies an epistemic mismatch. EU copyright law relies on qualitative standards interpreted contextually by human adjudicators. LLM development relies on quantitative metrics optimised by automated pipelines. Each side has a coherent internal logic. Together, they currently seem to fail to interlock (Lucchi and Hunter, 2025; Borhi et al., 2025; Quintais, 2025; Chen et al., 2024; Wei et al., 2024).

3.1 Qualitative Legal Standards

Copyright rights and exceptions are evaluated holistically. Originality depends on the author's own intellectual creation (Court of Justice of the European Union, 2009, 2019; European Parliament and Council, 2001). Substantial similarity is a totality-of-circumstances judgement that weighs expressive overlap, selection, and arrangement, and the overall impression, often alongside market substitution (Lucchi and Hunter, 2025). "Lawful access" under the TDM exception is contextual: it may depend on the interplay between statutory exceptions, licence terms, and ToS, and on whether rightsholders reserved rights "in an appropriate manner" (Quintais, 2025; Hamburg District Court, 2024; European Parliament and Council, 2019). These assessments are qualitative, fact-sensitive, and resolved in retrospect by courts.

3.2 Quantitative Technical Safeguards

LLM pipelines are governed by measurable proxies. Training optimises loss functions; filtering uses heuristics for de-duplication and quality; evaluation uses automatic metrics (e.g., ROUGE-L, BLEU, cosine similarity, edit distance, LCS, forget quality) and black-box probes for memorisation (Chen et al., 2024; Wei et al., 2024; Shi et al., 2024). Post-training safeguards such as machine unlearning, decoding filters, refusal policies are validated on benchmarks and scorecards (Wei et al., 2024; Maini et al., 2024; Shi et al., 2024). These instruments provide scalars and thresholds that can be embedded into CI/CD and governance tooling at scale.

3.3 Why the Edges Do Not Meet

The problem is not necessarily that the technical metrics are poor. The reason is that they answer different questions than the law asks.

Measuring the wrong thing. High ROUGE-L or cosine similarity may reveal overlap, but low scores do not certify the absence of substantial similarity. A passage can appropriate the selection, arrangement, or style of a work without triggering n-gram or embedding thresholds (Chen et al., 2024; Chun, 2024). In contrast, enforcement turns on market effects and expressive appropriation, which are not captured by token-level comparisons.

Certifying the impossible negative. Developers generally cannot conclusively prove that a given work was not included in training from opaque web scrapes, nor that a model will not produce infringing output under adversarial prompting (Borhi et al., 2025; Wei et al., 2024). Black-box probes and dataset summaries provide evidence, but cannot convert statistical uncertainty into the level of legal certainty the law typically demands.

Exceptions resist full automation. Whether TDM exceptions apply largely depends on lawful access and opt-outs expressed "in an appropriate manner". These depend on domains of provenance, licencing, and contract interpretation, that are inadequately addressed by relying solely on post-hoc statistical analyses (Quintais, 2025).

Benchmarks are not doctrines and typically only partially align with them. Unlearning benchmarks demonstrate reduced verbatim recall, but courts assess broader categories including derivative works, character and plot appropriation, and stylistic mimicry (Maini et al., 2024; Chun, 2024). Optimising to today's benchmarks can still leave tomorrow's legal requirements unmet.

3.4 Structural Non-Compliance

Even if developers minimise memorisation, adopt similarity thresholds, and publish high-level dataset summaries, they cannot reliably demonstrate conformance with qualitative legal standards. Conversely, even if rightsholders reserve rights and seek transparency, they cannot audit training at scale or map legal claims to technical artefacts. In short, current metrics do not prove compliance by themselves, and qualitative standards, without machine-actionable specifications, are difficult to apply at scale. This necessitates a new approach grounded in co-design.

3.5 Toward Co-Designed Frameworks

The path forward is neither to abandon metrics nor to dilute legal standards, but to co-design them into a coherent legal-technical compliance framework.

Law must shape metrics. Legal standards would benefit from translation into machine-actionable requirements: standardised opt-out schemas with clear precedence rules, provenance attestation formats tied to specific verification procedures, and disclosure templates that define "sufficiently detailed" in operational terms. This requires moving from aspirational principles to implementable specifications that pipelines can execute and auditors can verify.

Metrics should inform law. Legal doctrine could evolve to recognise families of technical evidence as meaningful for compliance determinations. Provenance graphs, certified unlearning bounds, adversarial robustness profiles, and synthetic data attestations should inform safe harbours and liability assessments. This requires courts and regulators to accept verifiable, reproducible technical evidence rather than demanding unattainable certainties about training data or future outputs.

Without this bidirectional translation, law and technology will remain orthogonal: retrospective, qualitative adjudication cannot govern prospective, quantitative systems at scale. Co-designed frameworks offer a credible path to making copyright compliance both technically tractable and legally meaningful in the age of LLMs.

4 Technical Mitigation: Promise and Limits of Current Safeguards

Technical safeguards for copyright compliance cluster around various pillars, such as provenance-first data governance, post-training removal via machine unlearning, and synthetic data generation. Each contributes to the reduction of risk. None, in its current form, appears sufficient, on its own, to establish enforceable compliance with EU legal standards.

4.1 Data Governance: Provenance, Licensing, and Opt-Outs

Input-side risk is primarily about reproduction. Provenance-first pipelines, such as ingestion workflows that restrict sources to lawful repositories, enforce machine-readable reservations, exclude shadow libraries, produce provenance manifests (source, hash, time, licence/opt-out status), reduce that risk by constraining what enters training.

Copyright-clean corpora and licensing. Curated datasets built from public-domain works, permissive licences, or negotiated content access likely offers the strongest legal footing (Bommarito II et al., 2025). Combined with licence-aware ingestion (ToS, parsing, whitelist sourcing, and shadow-library exclusion), this is the most direct approach to reducing unlawful reproduction at scale (Keller, 2024; Dornis, 2025). The cost is coverage: performance in specialised domains can degrade without high-quality proprietary sources, and curation is expensive (Fan et al., 2025).

Machine-readable reservations at scale. Parsing *robots.txt*, meta tags, *JSON-LD*, and natural-language ToS is often brittle and incomplete without a common schema and registry (Hamann, 2024; Keller, 2024). Stopgap industry tools (e.g., content registries or proposed protocols like TDM-REP) help but lack universal adoption and legal force. Absent a standardised EU opt-out infrastructure, "respecting opt-outs" remains a best-effort exercise rather than auditable compliance (European Commission, DG CNECT, 2025).

Transparency and attestations. Training-data transparency reports and data-lineage graphs can make inclusion decisions reviewable beforehand and auditable retrospectively. But without templates tied to thresholds (e.g., what granularity is "sufficiently detailed"?) and sampling/assurance protocols, disclosures risk being too high-level for rightsholders and too invasive for developers (Buick, 2024; Warso and Gahntz, 2024).

In summary, provenance controls are necessary to manage reproduction risk at the input stage. While they cannot resolve the underlying legal ambiguity of "lawful access" (Section 2), they could provide evidentiary support for compliance determinations if legal standards were clarified. However, today's mechanisms (schemas, registries, attestations) remain under-specified legally and under-developed technically, limiting their utility for demonstrating compliance.

4.2 Machine Unlearning: Removal After the Fact

Output-side risk arises when memorised or substantially similar content is produced. Unlearning methods attempt to remove or suppress targeted knowledge post-training, a capability driven not only by copyright takedown demands but also by data protection mandates like the GDPR's "right to erasure" (Article 17). Concretely, methods reduce likelihood on targeted spans (logit suppression, KL-regularised updates), use self-distillation to avoid catastrophic forgetting, and deploy reversible inference-time overlays. Despite promising drops in verbatim recall, current methods remain vulnerable to paraphrase leakage, adversarial prompting, and scaling limits for target discovery and application (Russinovich and Salem, 2025; Jin et al., 2024; Dong et al., 2025; Vasilev et al., 2025; Bhaila et al., 2025; Ji et al., 2024; Wei et al., 2024).

What works today. Token- or span-level methods (e.g., surgical logit suppression and KL-guided "unmemorisation") sharply reduce verbatim recall of known copyrighted passages while preserving general utility (Russinovich and Salem, 2025; Jin et al., 2024). Sequential unlearning approaches handle stages takedowns (e.g., multiple books over time) (Yao et al., 2024). Reference-free/self-distillation variants improve stability relative to naive gradient-ascent forgetting (Dong et al., 2025; Vasilev et al., 2025). Inference-time control (e.g., learned soft prompts, logit-difference patches) offer reversible, low-overhead deployment (Bhaila et al., 2025; Ji et al., 2024).

What still breaks. Scalability: applying targeted unlearning across millions of potential passages is compute-intensive and requires reliable detection/targeting (Xu et al., 2025). Coverage: most methods address literal copying, but they provide limited protection against paraphrase, stylistic mimicry, plot/character appropriation, or latent template reuse (Wei et al., 2024; Maini et al., 2024; Chun, 2024). Robustness: many methods can be bypassed by adversarial prompts, role-play framing, or jailbreak decoding strategies (Wei et al., 2024; Shi et al., 2024). Guarantees: certified unlearning is nascent and largely developed under convex assumptions that do not hold for LLMs; most demonstrations remain empirical rather than provable (Chien et al., 2024).

The bottom line is that unlearning belongs in the toolkit for managing output-side infringement and takedown workflows, potentially addressing both copyright takedown obligations and GDPR erasure rights. However, current methods' vulnerability to adversarial prompts and paraphrase leakage means they cannot yet guarantee legal compliance. Unlearning needs adversarially robust variants, scalable targeting, and formal legal recognition before it can serve as reliable evidence of compliance.

4.3 Synthetic Data Generation: Compliance-by-Design or Risk Laundering?

Given the limitations of both provenance controls and unlearning methods, synthetic data generation (SDG) has emerged as a potential third pillar for copyright compliance, but its compliance value depends on how the data are produced, ranging from risk-reducing approaches that transform clearly lawful inputs to risk-laundering approaches that prompt a general LLM as a knowledge source, with the latter inheriting the teacher's uncertain provenance.

Type-1 SDG (Clean Chain). The model is used purely as a tool to transform clearly lawful inputs (e.g., public-domain text, from licenced tables/graphs to text, translations, structured-to-text descriptions) (GPT-NL, 2024). Every token in the synthetic output is traceable to a lawful source. This approaches compliance-by-design: provenance is explicit, TDM exceptions may be unnecessary, and auditing becomes more tractable. However, type-1 SDG does not eliminate the lawful access requirement, it only shifts it upstream to the seed data. Developers must still ensure that the source materials were themselves lawfully accessed under applicable copyright and contract law. The limits are quality and coverage—especially for niche domains—plus the cost of building/curating lawful seeds at scale.

Type-2 SDG (Dirty Chain). Outputs are generated from a general LLM as a knowledge source. This imports the unknown training provenance of the teacher model and can reintroduce verbatim fragments or close paraphrases (GPT-NL, 2024). It also risks "model collapse" when recursively training on synthetic outputs, degrading quality, and amplifying biases (Shumailov et al., 2024). From a compliance perspective, type-2 SDG tends to launder uncertainty rather than remove it: aggressive filtering helps, but cannot restore a clean-chain of provenance.

The bottom line is that type-1 SDG is a strong ingredient for input-side compliance strategies by enabling expansion within the boundaries of demonstrably lawful sources, although the lawful access requirement persists for seed data. Type-2 SDG should generally be treated as high-risk, as it inherits the teacher model's uncertain provenance rather than establishing an independent lawful basis.

4.4 Synthesis: Necessity-Insufficiency Gap

A credible technical posture for copyright compliance combines provenance infrastructure for input-side control (licence-aware ingestion, standardised opt-outs, data-lineage attestations); adversarially robust unlearning and inference-time defences for output-side control (targeted removal, robustness testing, reversible controls); and type-1 synthetic data to expand lawful coverage without importing copyright risk.

This technical triad is likely necessary, but, on its own, may remain insufficient to bridge the epistemic gap between technical implementation and legal proof. The following essential gaps remain at the intersection of law and technology:

- Legal Recognition: Technical artefacts (provenance attestations, unlearning reports, robustness profiles) lack formal status in compliance determinations or safe harbour provisions.

- Evaluation Misalignment: Existing benchmarks optimise for literal overlap rather than legally salient harms (substantial similarity, derivative works, market substitution effects).

- Adversarial Brittleness: Current defences are vulnerable to prompt engineering, jailbreaking, and distributional shifts that may trigger infringement despite clean training data.

4.5 Implications for Compliance Strategies

Technical safeguards can dramatically reduce copyright risk, but without legally defined, machine-readable standards upstream and legally meaningful evaluation downstream, they cannot prove compliance. The right objective is not a single silver bullet, but a co-designed pipeline: provenance-first ingestion, lawful expansion via type-1 SDG, and robust, auditable takedowns—embedded in standards that regulators can verify and rightsholders can rely on.

5 Path Forward: Policy and Research Directions

Bridging the legal-technical gap requires co-designed standards that are machine-actionable upstream and legally meaningful downstream. The direction below outlines plausible pathways for policymakers, developers, and researchers to make copyright compliance verifiable at scale. They are proposed as guidance rather than prescriptive steps.

5.1 Policy Directions

Standardise machine-readable opt-outs. Policymakers could consider adopting a harmonised EU schema with clear precedence over *robots.txt* (which was never created with large-scale AI scraping in mind) and natural-language ToS. Provide a conformance profile to enable reliable parsing in ingestion pipelines.

Clarify core terms and disclosures. Provide guidance clarifying the interpretation of ambiguous or "undefined" terms such as "lawful access" (relationship to paywalls, rate limits, ToS, and licences) and what qualifies as a "sufficiently detailed" training-data summary (coverage categories, licensing classes, time windows, acceptable aggregation, and sampling/attestation practices).

Explore centralised registries. Explore the feasibility of an EU-facing portal and API through which rightsholders can register opt-outs and licences, and developers can retrieve authoritative signals during crawling and ingestion.

Develop audit baselines. Develop baseline documentation templates and lightweight audit checklists covering data provenance controls, opt-out handling, takedown workflows, and disclosure practices.

Support interdisciplinary compliance labs. Support legal and technological collaborations to co-develop and pilot: legally informed evaluation tasks for substantial similarity, derivative works, and market-substitution proxies; copyright-specific adversarial robustness protocols and reporting formats; evidence standards (e.g., provenance attestations, unlearning reports, robustness scorecards) that could underpin presumptions or safe harbours.

5.2 Developer Practices

Build licence-aware ingestion. Implement terms/licence parsing, integrate machine-readable opt-out signals where available, exclude shadow libraries, and maintain provenance graphs (URLs, hashes, timestamps, licence/opt-out status, and exclusion rationales) for training artefacts.

Prefer clean-chain synthetic data. Use type-1 synthetic data (transformations of lawful sources, such as public-domain and licenced structured data) to extend coverage. Treat type-2 synthetic data as generally high-risk unless teacher provenance is demonstrably clean and outputs pass stringent deduplication and similarity filters.

Integrate unlearning and defences. Adopt span-level unmemorisation for known infringements, sequential unlearning for staged takedowns, and inference-time defences (refusal policies, decoding constraints). Establish internal red-teaming suites where feasible for copyright-specific jailbreaks and evaluation mitigation before release.

Document and attest. Publish training-data summaries aligned with legal templates once available. Maintain internal, signed provenance manifests and unlearning/robustness reports to support regulator and rightsholder inquiries.

5.3 Research Avenues

Benchmarks tied to doctrine. Move beyond a primary focus on literal overlap to tasks and metrics that approximate legally salient harms: substantial similarity (stylometry, event/character graphs, selection-and-arrangement), derivative works and style appropriation (style-transform detection; architectural-pattern similarity), and market substitution proxies (simulation-based indicators).

Multi-domain unlearning. Develop methods that jointly handle copyright and GDPR erasure, with parameter localisation, robustness to adversarial prompts, and minimal collateral damage to unrelated knowledge.

Certified and scalable forgetting. Extend certified unlearning concepts to non-convex LLMs. Design efficient targeting and verification at corpus scale. Provide reversibility and audit trails suitable for takedown workflows.

Robustness as evidence. Standardise red-teaming protocols focused on copyright extraction and paraphrase leakage, reporting formats and acceptance thresholds that correlate with legal risk.

Provenance instrumentation. Advance scalable provenance capture (lineage graphs, content hashes, licence/opt-out metadata) and sampling-based assurance that can be independently verified.

6 Conclusion

This paper has suggested that a persistent epistemic gap exists between EU copyright law and the practical realities of LLM development, creating a risk of structural non-compliance. We have shown that the law's qualitative, retrospective standards of originality and substantial similarity are fundamentally misaligned with the quantitative, proactive controls that govern machine-scale pipelines. Consequently, existing legal instruments like the TDM exceptions and GPAI transparency duties are ill-matched to the challenges of petabyte-scale training, while technical safeguards such as provenance-first governance, MU, and SDG remain insufficient without legally-specified standards and meaningful evaluation.

The path forward is not a single technical fix or legal decree, but the development of a coherent, auditable ecosystem built on legal-technical co-design. This requires translating legal duties into machine-actionable specifications upstream, while enabling technical evidence to be legally meaningful downstream. Such an ecosystem would be founded on provenance-first data ingestion, lawfully expanded via clean-chain synthetic data, and protected by robust, verifiable corrective techniques, such as machine unlearning. This integration could make compliance testable, proactive controls auditable, and retrospective adjudication reliant on verifiable artefacts rather than unattainable certainties. Achieving this vision, however, will require coordination between legal and technical communities that may prove difficult in practice, and our technical recommendations await empirical validation at scale, with implementation costs likely varying significantly across organisational contexts.

Ultimately, the EU faces a strategic choice: work toward enforceable compliance frameworks grounded in this integrated approach, or risk gradual erosion of copyright relevance in the age of generative AI. Our analysis focuses specifically on EU copyright law and may not generalise to jurisdictions with fundamentally different frameworks, such as the US fair use doctrine. Moreover, whilst we outline policy directions toward standardised schemas and registries, the political economy of implementation involves stakeholder interests beyond our scope. Our focus on copyright compliance necessarily brackets other relevant legal frameworks such as data protection under GDPR, competition law, and sector-specific regulations that may interact with the proposed solutions in ways requiring further research. Nevertheless, by fostering policy, development, and research that bridge the legal-technical gap, and by remaining responsive to the rapid evolution of both legal developments and technical capabilities, the EU can set the global standard for responsible AI: protecting creators' rights while enabling innovation at machine-scale.

Acknowledgments

We thank the Contractuo team for constructive feedback and discussions. This research was conducted during an internship at Contractuo and as part of a MSc thesis at Maastricht University. This work is part of NWO-TDCC project (ICT.001.TDCC.014 with budget number 20656). The views expressed are those of the authors and do not necessarily reflect those of Contractuo or Maastricht University.

Ethical Considerations

This position paper addresses the ethical implications of large-scale training on copyrighted content without explicit consent. Our analysis aims to protect creator rights whilst enabling responsible AI development.

Potential harms. The current misalignment between legal frameworks and LLM technology creates several ethical concerns: systematic copyright infringement at scale, inability of creators to meaningfully consent to or opt out of training, and economic displacement without compensation. These harms disproportionately affect individual creators and smaller rightsholders who lack resources to pursue enforcement.

Broader impact. Our proposed co-design framework aims to mitigate these harms by making copyright compliance technically tractable and legally meaningful. However, implementation of our recommendations could impose costs on AI developers and potentially limit access to certain training data, affecting model performance and innovation velocity. We argue these trade-offs may be necessary to preserve creator rights and maintain public trust in AI systems.

References

Karuna Bhaila, Minh-Hao Van, and Xintao Wu. 2025. Soft prompting for unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4046–4056, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael J. Bommarito II, Jillian Bommarito, and Daniel Martin Katz. 2025. The kl3m data project: Copyright-clean training resources for large language models. *arXiv*.

Maurizio Borhi, Bryan Khan, Anna Arnaudo, Riccardo Raso, Marco Ricolfi, Antonio Vetro, Riccardo Coppola, Antoine Aubert, Ziga Drobnic, Stephan Edelbroich, Chikemka Abuchi-Ogbonda, and Raffaele Darroch. 2025. The development of generative artificial intelligence from a copyright perspective. Technical report, European Union Intellectual Property Office (EUIPO). Study.

Adam Buick. 2024. Copyright and ai training data—transparency to the rescue? *Journal of Intellectual Property Law & Practice*, 20(3).

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. 2024. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15134–15158, Miami, Florida, USA. Association for Computational Linguistics.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. 2024. Certified machine unlearning via noisy stochastic gradient descent. *arXiv*.

Jon Chun. 2024. AIStorySimilarity: Quantifying story similarity using narrative for search, IP infringement, and guided creativity. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 161–177, Miami, FL, USA. Association for Computational Linguistics.

A. Feder Cooper, Aaron Gokaslan, Ahmed Ahmed, Amy B. Cyphert, Christopher De Sa, Mark A. Lemley, Daniel E. Ho, and Percy Liang. 2025. Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv*, 2505.12546. V2.

Court of Justice of the European Union. 2009. Case c-5/08, infopaq international a/s v danske dagblades forening. EUR-Lex. ECLI:EU:C:2009:465.

Court of Justice of the European Union. 2019. Case c-683/17, cofemel – sociedade de vestuário sa v g-star raw cv. EUR-Lex / CURIA. ECLI:EU:C:2019:721.

Artha Dermawan. 2024. Text and data mining exceptions in the development of generative ai models: What the eu member states could learn from the japanese "nonenjoyment" purposes? *The Journal of World Intellectual Property*, 27(1):44–68.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2025. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840, Albuquerque, New Mexico. Association for Computational Linguistics.

Tim W. Dornis. 2025. Generative AI, Reproductions Inside the Model, and the Making Available to the Public. *IIC International Review of Intellectual Property and Competition Law*, 56(5).

European Commission, DG CNECT. 2025. Tender specifications: Feasibility study on a registry for tdm opt-outs. Tender specifications EC-CNECT/2025/OP/0002, European Commission, Brussels.

European Parliament and Council. 2001. Directive 2001/29/ec of the european parliament and of the council of 22 may 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal of the European Union*.

European Parliament and Council. 2019. Directive (eu) 2019/790 of the european parliament and of the council of 17 april 2019 on copyright and related rights in the digital single market and amending directives 96/9/ec and 2001/29/ec. *Official Journal of the European Union*.

European Parliament and Council. 2024. Regulation (eu) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). *Official Journal of the European Union*.

Dongyang Fan, Vinko Sabolčec, Matin Ansaripour, Ayush Kumar Tarun, Martin Jaggi, Antoine Bosselut, and Imanol Schlag. 2025. Can performant llms be ethical? quantifying the impact of web crawling opt-outs. *arXiv*.

GPT-NL. 2024. Synthetische data. GPT-NL — Nieuws. Web article (in Dutch), initiative website.

Hanjo Hamann. 2024. Artificial intelligence and the law of machine-readability: A review of human-to-machine communication protocols and their (in)compatibility with article 4(3) of the copyright dsm directive. *Journal of Intellectual Property, Information Technology, and Electronic Commerce Law*, 15(2).

Hamburg District Court. 2024. Germany — hamburg district court, 310 o 227/23, robert kneschke v LAION e.v. WIPO Lex. Landgericht Hamburg, 310 O 227/23 (2024-09-27).

Štěpánka Havlíková. 2025. Technical challenges of rightsholders' opt-out from gen ai training after robert kneschke v. laion. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 16(1).

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *arXiv*.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*. Datasets and Benchmarks Track, poster.

Paul Keller. 2024. Considerations for opt-out compliance policies by ai model developers. Technical report, Open Future.

Nicola Lucchi and Serra Hunter. 2025. Generative ai and copyright: Training, creation, regulation. Study PE 774.095, European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs (JURI), Brussels. Requested by the Committee on Legal Affairs (JURI).

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv*.

Thomas Margoni and Martin Kretschmer. 2022. A deeper look into the eu text and data mining exceptions: Harmonisation, data ownership, and the future of technology. *GRUR International*, 71(8).

João Pedro Quintais. 2025. Generative ai, copyright and the ai act. *Computer Law & Security Review*, 56.

Mark Russinovich and Ahmed Salem. 2025. Obliviate: Efficient unmemorization for protecting intellectual property in large language models. *arXiv*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Stefan Vasilev, Christian Herold, Baohao Liao, Seyyed Hadi Hashemi, Shahram Khadivi, and Christof Monz. 2025. Unilogit: Robust machine unlearning for LLMs using uniform-target self-distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22453–22472, Vienna, Austria. Association for Computational Linguistics.

Zuzanna Warso and Maximilian Gahntz. 2024. Advancing training data transparency in the eu ai act. Open Future Blog. Blog post.

Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. 2024. Evaluating copyright takedown methods for language models. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*. Curran Associates, Inc.

Tianyang Xu, Xiaoze Liu, Feijie Wu, Xiaoqian Wang, and Jing Gao. 2025. Suv: Scalable large language model copyright compliance with regularized selective unlearning. In *Proceedings of the Conference on Language Modeling (COLM 2025)*. COLM 2025.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.