Training Large Language Models on Web-Scraped Personal Data

Legal, Ethical, and Regulatory Implications

A dissertation submitted in partial fulfilment of the requirements

for the degree of BSc (Hons) Computer Science

Amin Wafi

Student ID: 100675781

Supervisor: Stavros Roupakas

Module: 6CM995 Individual Project

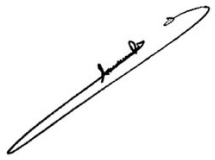
University of Derby

May 2026

Declaration of Originality

I, Amin Wafi, declare that this dissertation is my own work and has not been submitted, in whole or in part, for any other academic award. Where I have drawn on the work of other authors, I have indicated this clearly in the text and in the consolidated reference list at the end of the document. The analytical positions defended in this dissertation are my own, as are the doctrinal commitments I make and the conclusions I draw from the evidence and arguments presented.

I confirm that this work has been prepared in accordance with the University of Derby's academic integrity policy. Where I have used digital tools to support drafting, structuring, and editing, the analytical content, source selection, and the position defended remain my own intellectual responsibility.



Signed: _____

Date: 26 May 2026

Student ID: 100675781

Acknowledgements

I am grateful to my supervisor, Stavros Roupakas, for guidance and feedback throughout this project, and for the patience required to let an argument find its shape across several drafts. I would also like to thank my family for their support over the course of this dissertation, without which the work would not have been finished. Any errors and overstatements that remain are mine alone.

Abstract

Large language models are now trained on hundreds of billions of pages of web-scraped content, much of which contains personal data that was never offered to the developers using it. This dissertation argues that the GDPR cannot meaningfully govern that activity. The failure is not contingent on regulatory craft or developer good faith, and it cannot be repaired through clearer guidance or better documentation. It reflects a structural mismatch between a regulatory framework built around bounded, purpose-specific processing relationships and a form of data use that has none of those properties.

The argument develops through a desk-based, integrative methodology that combines doctrinal analysis, empirical research on memorisation and dataset composition, and normative work on fundamental rights and collective harm. The central GDPR obligations cannot be satisfied at the scale at which frontier LLMs are now built; the strict reading of Article 9(2)(e) following GC and Others v CNIL is incompatible with web-scale corpora that unavoidably contain special-category data. The EU Charter as developed in Digital Rights Ireland and La Quadrature du Net prohibits the kind of indiscriminate processing that web-scraped LLM training amounts to, a question the regulatory debate has barely engaged. The most prominent enforcement effort to date, the Italian Garante's €15 million fine against OpenAI, was annulled by the Court of Rome in March 2026, leaving Europe with no completed enforcement decision to point to.

The contribution is to connect three literatures usually kept apart and to test the existing regulatory responses (the EDPB framework, the CNIL's June 2025 guidance, the AI Act's Article 53 template) against a standard developed from Charter case law, contextual integrity theory, and collective rights doctrine. The position is that the regulatory architecture has to change to fit the activity, the activity has to change to fit the architecture, or some combination of the two. What cannot continue is a regime designed for one form of processing being asked to govern another, producing dense documentation and modest disclosure without the protection the regime was designed to deliver.

Keywords: GDPR; large language models; web scraping; data protection; EU Charter of Fundamental Rights; AI Act; collective rights; contextual integrity.

Table of Contents

Chapter 1. Introduction

1.1 The problem

Frontier large language models are trained on vast quantities of personal data scraped from the open web. The corpora behind systems like GPT-4, Claude, and Gemini contain hundreds of billions of pages drawn from Common Crawl and supplemented by social media content, professional networks, government records, patient support forums, personal blogs, and the everyday textual residue of contemporary online life. Almost none of this data was provided to the developers who use it. It was collected by automated scrapers from publicly accessible web pages, aggregated into training datasets, and used to build commercial AI systems whose deployment now affects hundreds of millions of users.

The legal question this raises is straightforward in the asking and surprisingly difficult in the answering. Does this kind of processing comply with the General Data Protection Regulation (European Parliament and Council, 2016), the EU instrument that governs the processing of personal data of individuals within the EU and EEA? Regulators have engaged the question with increasing seriousness since the public release of ChatGPT in November 2022 made the scale of the activity commercially visible. The Italian Garante imposed a temporary block on ChatGPT in March 2023 and a €15 million fine in November 2024. The European Data Protection Board's ChatGPT Taskforce reported in May 2024 (EDPB, 2024a). EDPB Opinion 28/2024 was adopted in December that year (EDPB, 2024b). The CNIL issued operational guidance in June 2025. The EU AI Act's transparency obligations for general-purpose AI models took effect in August 2025. The pace of regulatory activity has been considerable.

The pace of regulatory clarity has been less so. The EDPB has insisted that the GDPR applies in full to LLM training while acknowledging that compliance is practically difficult (EDPB, 2024a, 2024b). The Garante's fine was annulled by the Court of Rome in March 2026, six weeks before this dissertation was submitted. The CNIL's guidance assumes that legitimate interest can be made to work as a legal basis without explaining how the underlying balancing test can be conducted at scale. The AI Act and the GDPR operate as parallel regimes with thin coordination. Across Europe, regulators understand the technical and legal questions in considerable detail, and yet not a single sustained enforcement action against a frontier-model developer has produced an outcome the courts have allowed to stand.

My argument is that this is not a coincidence, and that explaining it requires looking at the regulatory framework rather than at the regulators applying it.

1.2 Thesis

The position I defend is that GDPR compliance is not contingently but structurally unavailable to frontier-scale LLM development on web-scraped personal data. The framework's difficulties in this domain are not the kind that better engineering, more careful documentation, or sharper guidance can repair. They reflect a mismatch between the regulatory architecture and the activity it is being asked to regulate, and the mismatch shows up at three connected levels.

At the doctrinal level, the GDPR's central obligations cannot, on a defensible reading, be satisfied by the activity. Article 6's lawful basis, the transparency obligations under Articles 13 and 14, the special-category prohibition in Article 9, the data subject rights of Articles 15 to 22, and the design-level requirements of Article 5 read with Article 25 each break down at points the analysis identifies. The breakdowns are not random. They reflect the fact that the framework was built for processing relationships that are bounded, purpose-specific, and individually addressable, and the activity it is being asked to govern has none of those properties.

At the constitutional and ethical level, the protective values that the GDPR is meant to deliver extend beyond the GDPR's own scheme. Articles 7 and 8 of the EU Charter, as developed in Digital Rights Ireland, Schrems II, and La Quadrature du Net, prohibit indiscriminate processing of personal data and require demonstrated necessity, proportionality, and procedural safeguards adequate to the seriousness of the interference. Web-scraped LLM training instantiates exactly the doctrinal pattern those cases identified as Charter-incompatible, and the question has barely been engaged in current practice. The harms involved are also collective and diffuse rather than individually particularised, and an individual-rights framework is not built to detect or remedy them.

At the level of governance, current responses do not address these structural problems. The EDPB's framework is careful but operates within the existing architecture rather than questioning it (EDPB, 2024a, 2024b). The Garante's enforcement effort has not produced sustained outcomes. The CNIL's guidance is operationally specific but does not engage the deeper issues. The AI Act's transparency template provides corpus-level disclosure without

addressing individual-level privacy concerns. The UK position is essentially the EU position with the enforcement layer removed.

The central claim is that these three levels of failure are connected, and that connecting them changes the policy debate. The framing in which compliance is hard but ultimately achievable through better regulatory craft does not match the structural picture the analysis reveals. Reform would have to operate at the level of regulatory architecture rather than at the level of compliance mechanisms.

1.3 Research questions

Three research questions structure the dissertation. The answers are developed across the analytical chapters and synthesised in Chapter 7.

RQ1 asks whether GDPR compliance is practically unavailable to large-scale LLM development, in the sense that the legitimate interests balancing test under Article 6(1)(f) and the transparency obligations under Articles 13 and 14 presuppose conditions that web scraping at scale makes impossible to satisfy. The answer the dissertation defends is yes, on the doctrinal reading I argue for, with the qualification that the answer assumes the EDPB's functional reading of "personal data" prevails over the Hamburg DPA's narrower technical reading.

RQ2 asks whether the individual-rights architecture of the GDPR produces a misalignment with the collective and systemic character of the harms LLM training generates, and what this means for whether legal compliance is even the right standard to apply. The answer the dissertation defends is that the misalignment is real, that the values being un-honoured extend significantly beyond the GDPR's own scheme into the constitutional and ethical domains, and that legal compliance, even if achieved, would not be sufficient.

RQ3 asks what governance reforms, feasible within the existing GDPR and AI Act framework, could address the root causes of the compliance problem and the power asymmetry between AI developers and the people whose data they use. The answer is the substance of Chapter 7, which sets out the design requirements any adequate reform would have to meet and develops a data trusts proposal as one worked version of what meeting those requirements would look like.

1.4 Structure

The dissertation is built across seven chapters. Chapter 2 reviews the relevant literature across five thematic areas. Chapter 3 sets out the methodology. Chapter 4 develops the empirical picture the doctrinal chapter then relies on. Chapter 5 conducts the doctrinal GDPR analysis. Chapter 6 develops the fundamental rights, ethical, and governance assessment. Chapter 7 synthesises the analysis, answers the research questions, and sets out the design requirements that follow.

1.5 Scope and positionality

Some boundary-setting is appropriate at the outset. The dissertation focuses on GDPR-governed processing in the EU. The UK is engaged in Chapter 6 because the institutional context of the dissertation is a UK university, but the comparison is not exhaustive. The US, Chinese, and other major regulatory environments are not engaged in any depth. The activity examined is frontier-scale foundation-model training on web-scraped personal data; smaller-scale, consent-based, or domain-specific AI training falls outside the analysis except where it appears as a possible alternative the GDPR could in principle accommodate. Within the GDPR itself, I focus on the obligations most directly relevant to LLM training. Not every Article is engaged.

My own position is that of a final-year Computer Science student with some technical familiarity with how machine learning systems are built and an analytical interest in the regulatory questions they raise. The analysis is informed by both. I do not claim the doctrinal authority of a legal practitioner, and the dissertation does not contribute new findings to the technical literature on memorisation or dataset construction. What I do claim is that the integrative work of connecting these literatures is worth doing, and that doing it carefully reveals features of the regulatory question that are harder to see when each literature is read on its own.

The dissertation defends a position rather than presenting both sides neutrally. That position is that GDPR compliance is structurally unavailable to frontier-scale LLM training, and that the reforms worth pursuing are the ones that engage the constitutional and collective dimensions the analysis identifies rather than the ones that try to refine existing compliance mechanisms. There are real objections to this. The procedural reading of Article 5 has

substantial scholarly support, the public-private distinction in Charter case law might do more work than I have allowed, and methodological individualists have principled reasons to resist collective-rights proposals. The chapters say where the objection persuaded me and where it did not. They also flag which doctrinal commitments would need adjusting if the objection ultimately prevails. The dissertation is a defended position, not a settled reading.

Chapter 2. Literature Review

This chapter reviews the literature across five thematic areas relevant to the question of whether GDPR compliance is practically available to frontier-scale LLM development on web-scraped personal data. The review establishes the empirical reality of training data composition, the legal position under the GDPR, the fundamental rights framework that sits behind data protection law, the ethical critique of consent and power asymmetry in commercial AI development, and the adequacy of the regulatory responses that have emerged since the public release of ChatGPT in late 2022. The chapter ends with a synthesis identifying why governance keeps failing in this area and what the analytical chapters will need to show to support the dissertation's thesis.

2.1 Introduction

This review is organised around a specific and largely unresolved tension between how large language models are actually built and what EU data protection law requires. The GDPR was designed around a clearly bounded relationship: a controller collects data from an identified or identifiable person, explains the purpose at the point of collection, and is legally constrained to that stated purpose. LLM development does not work like that. Building these systems involves pulling text from billions of web pages, most of it written by people who have no idea their words are being used for this purpose and who have no realistic way to object even if they did know. That incompatibility is structural, not incidental, and the regulatory literature has been slower than it ought to be in acknowledging it plainly.

Three research questions organise this review. First, what does the empirical evidence establish about the nature and extent of personal data in web-scraped LLM training corpora, and what are the legal consequences of those findings? Second, is GDPR compliance, particularly through the legitimate interests basis under Article 6(1)(f), realistically achievable at the scale and operational character of LLM training, or does the architecture of that training make certain compliance requirements structurally unavailable? Third, is the current governance response, including regulatory guidance, technical mitigations, and the EU AI Act, adequate to address the harms produced, and if not, what structural changes to enforcement design would be required?

Two analytical frameworks run through this review alongside the doctrinal legal analysis. The first is Nissenbaum's (2019) contextual integrity argument, which holds that privacy violations occur when information moves in ways that break the norms of the context in which it was originally shared. That framework helps describe what is fundamentally wrong with web scraping as a data collection method, something the GDPR's unlawful processing concept does not fully capture, but it also has significant limitations as a prescriptive tool that are engaged with critically rather than set aside. The second is Mantelero's (2022) case for collective rights, which argues that individual rights mechanisms are structurally inadequate when harms are diffuse, aggregate, and group-level rather than personal. That argument is central to the synthesis this review develops: the consistent failure of governance in this space is explained not by regulatory ignorance but by an architectural mismatch between individual rights enforcement design and harms that are fundamentally collective in character.

The review moves through what LLM training datasets actually contain and why that matters legally, whether GDPR compliance is realistically achievable at web-scraping scale (with particular focus on the legitimate interests basis and the transparency obligations), what the EU Charter of Fundamental Rights adds beyond the protections available under secondary law, and how contextual integrity and collective rights theory illuminate the gaps that legal compliance alone leaves. The ethical stakes of power asymmetry, consent, and industry self-regulation are then examined, before the question of how adequate current governance responses actually are, including the EU AI Act. The closing section draws those threads together to develop the central argument about why governance has consistently failed and what would genuinely need to change.

2.2 The empirical reality of training data

A significant portion of industry and policy discussion frames the privacy problem in AI training data as essentially a contamination issue: useful text gets collected, some personal information is inadvertently included, and improved filtering can progressively reduce that over time. The empirical research does not support that framing, and it is worth being precise about why, because the distinction matters for the legal analysis that follows.

Hong et al. (2025) conducted a systematic audit of a prominent web-scraped machine learning dataset and found significant volumes of personally identifiable information still present after prior sanitisation had already been applied. This included medical details,

financial records, and data concerning minors. That finding matters not merely because one dataset exhibited privacy deficiencies, but because it reflects something structural about the open web itself. People share personal information online because that is how the web functions: they write about themselves, publish professional details, discuss experiences, and document their lives. Any dataset constructed from the open web at scale will therefore contain personal data not because filtering is imperfect, but because the raw material contains personal data by design. The EDPS TechSonar report (EDPS, 2023) reaches the same conclusion from a regulatory standpoint, observing that the architectural requirements of state-of-the-art LLMs make privacy exposure at the collection stage close to inevitable rather than contingent on any particular design choice.

It is worth noting that Hong et al. (2025) is, at the time of writing, an unreviewed preprint. The weight given to it here reflects the consistency of its findings with the broader direction of empirical privacy research on machine learning datasets, rather than reliance on it as a uniquely authoritative source. The structural claim, that web-scraped datasets will contain personal data as a predictable consequence of how the web operates, is independently supported by the EDPS (2023) and does not depend on any single empirical study.

The downstream implications for deployed models are equally significant and raise a legal question the literature has been insufficiently precise about. Carlini et al. (2021) demonstrated that under certain prompting conditions, LLMs can reproduce verbatim content from training data, including private personal information that was present in that data. Carlini et al. (2023) extended this finding to show that memorisation scales with model size: the most capable, commercially deployed models are also the most prone to privacy disclosure. This creates a legal consequence that current regulatory guidance has not fully engaged with. If personal data processed without a valid lawful basis has been memorised into the weights of a deployed model, the ongoing availability and use of that model may constitute a continuing GDPR breach rather than merely a historical processing failure, because the personal data remains encoded in the system and capable of reproduction. The question of whether deployed model weights containing memorised personal data constitute "processing" within the meaning of Article 4(2) GDPR is unresolved in current regulatory guidance and represents a practically significant gap.

The standard industry response to memorisation risk is to rely on anonymisation as a legal route out of data protection obligations. The problem is that anonymisation only serves as an

exit from the GDPR framework if it can reliably and irreversibly break the link between a trained model and the personal data used to train it. The evidence from Carlini et al. (2023), with the qualifications about generalisation from research conditions to deployed-system rates noted in Chapter 4, strongly indicates that at the performance levels at which frontier models currently operate, that standard cannot be met: differential privacy techniques sufficient to prevent memorisation-based disclosure impose accuracy costs that commercially deployed models do not accept. Solove (2021) adds a further dimension: notice-and-consent frameworks are structurally designed for discrete, bounded processing relationships and place obligations on data subjects that are impossible to discharge when data is collected from millions of unknown individuals at once. Both of the main GDPR compliance routes available at scale, anonymisation and notice, are practically unavailable under the conditions of LLM training at the current frontier.

2.3 GDPR compliance at AI scale

The central GDPR question for LLM training is lawful basis. Most major developers have relied, at least implicitly, on Article 6(1)(f): legitimate interests. EDPB Opinion 28/2024 (EDPB, 2024b) accepted that legitimate interests can in principle apply to AI model development, subject to the three-part test: the interest must be legitimate; processing must be necessary and proportionate; and the interests or fundamental rights of data subjects must not override those of the controller. The third condition is where the serious analytical problems begin.

The legitimate interests balancing test requires a genuine and documented weighing of the controller's commercial or technical interests against those of identifiable data subjects. At LLM training scale, those data subjects are not identifiable to the controller. They are unknown individuals whose data has been collected en masse from billions of documents without any direct contact with, or even awareness of, the people those documents concern. A genuinely rigorous balancing exercise requires that the controller know something substantive about the people on the other side of the balance: what data they contributed, what processing they would reasonably anticipate, and what the impact of the proposed use is likely to be on people in comparable positions. Where those people are entirely unknown and their data unexamined, that exercise cannot be conducted with any real substance. The EDPB (2024b) simultaneously requires rigorous conduct of this test and acknowledges the practical difficulty of meeting transparency obligations when processing data from millions of

unknown individuals, without resolving the tension between those two positions. Guidance that correctly identifies a compliance gap without providing a pathway through it generates legal uncertainty rather than substantive protection.

This tension is not merely a gap in guidance but reflects a structural incompatibility that Kamarinou, Millard and Singh (2017) identified in principle before LLM development had reached its current scale. They argued that the combination of purpose limitation, data minimisation, and transparency obligations under the GDPR creates systematic friction with the data-maximisation logic of modern AI systems, and that procedural compliance mechanisms, including documentation, impact assessments, and consent mechanisms, cannot resolve that friction because it is embedded in the design of both the law and the technology. That diagnosis has become considerably more acute as model scale has increased.

The CNIL (2025a) and ICO (2023) have each tried to operationalise the legitimate interests framework through concrete guidance requiring documented risk assessments, opt-out mechanisms, and safeguards proportionate to the processing. Both represent genuine regulatory effort, and the CNIL's guidance in particular is more practically specific than earlier EDPB material. Both sets of guidance share a common limitation, however: they require rigorous documentation of a balancing exercise that, for the reasons described above, cannot actually be conducted properly where data subjects are entirely unknown. Documenting an assessment that cannot be substantively carried out does not cure its deficiency; it produces a paper trail for a process whose foundations are missing.

The transparency obligations under Articles 13 and 14 present an independent and arguably more fundamental problem. Article 14, which governs collection from sources other than the data subject, requires equivalent privacy information within one month. The EDPB (2024b) insists these obligations apply in full to LLM training data collection while simultaneously acknowledging that they are practically impossible to meet at scale. Unlike the balancing test, there is no obvious procedural workaround: both Articles require individual-level notice, and the data subjects are not individually identifiable. Solove (2021) argues that this kind of impossible obligation reflects a deeper design failure: notice-and-consent frameworks were built for the bounded processing relationships that characterised data collection in the late twentieth century, and mass collection from diffuse public sources has made the foundational assumptions of that design obsolete.

One important limitation should be acknowledged at this stage. The regulatory position across the GDPR, EDPB Opinion 28/2024 (EDPB, 2024b), and the CNIL and ICO guidance is actively developing, and some apparent contradictions in the current guidance may be resolved by material published after this review was completed. The analysis reflects the best available reading of the regulatory position as of early 2026, and conclusions about irresolvable tension should be understood as provisional rather than permanent.

2.4 Fundamental rights and what compliance does not cover

Even if full GDPR compliance were achievable in technical terms, that would not conclude the normative analysis. The GDPR is secondary EU law and must be interpreted consistently with, and in light of, Articles 7 and 8 of the EU Charter of Fundamental Rights (European Union, 2012), which protect the right to respect for private life and the right to protection of personal data as fundamental rights at constitutional level. The significance of this for the present analysis is that GDPR compliance which systematically and substantially undercuts the autonomy and informational self-determination values those Charter articles protect is not merely ethically questionable; it may be constitutionally inadequate, regardless of whether each individual Article of the GDPR has been technically satisfied.

Mantelero (2022) provides the most systematic development of this concern. He argues that compliance-focused approaches to AI governance risk collapsing fundamental rights into procedural checklists: satisfying the formal requirements of a data protection impact assessment, documenting a lawful basis, and providing privacy notices comes to be treated as equivalent to protecting the autonomy, dignity, and self-determination that Articles 7 and 8 of the Charter are actually intended to secure. In the context of LLM training on web-scraped data, the gap between those two things is wide. The extraction of personal information from hundreds of millions of people who are entirely unaware of the process involves a harm to autonomy and informational self-determination that procedural compliance metrics, which were designed to regulate bounded, identifiable processing, were never built to capture.

Nissenbaum's (2019) contextual integrity framework offers a useful complementary lens for making the nature of that harm more precise. Privacy violations, on this account, do not occur merely when personal data is processed without a valid lawful basis; they occur whenever information flows in ways that break the contextual norms governing the setting in which it was originally produced and shared. A personal blog post written for a self-selected audience,

a professional profile published in a specific occupational context, or a message board contribution directed at a particular community each carries implicit norms about what flows of that information would and would not be appropriate. When that content is extracted and incorporated into a commercial AI training pipeline, those contextual norms are violated regardless of whether any GDPR article has technically been breached, because the flow that has occurred is fundamentally different from any flow the original context could have rendered appropriate.

The contextual integrity framework is analytically valuable here, but its limitations must be engaged with seriously. As a diagnostic tool it is compelling, often clearer than GDPR analysis because it does not depend on establishing statutory "personal data" or the absence of a lawful basis. As a prescriptive framework its utility is weaker. It does not resolve conflicts between competing contextual norms, it struggles when the technology being evaluated is itself reshaping the contexts it depends on, and where an LLM's outputs synthesise content drawn from many original contexts simultaneously, the framework cannot straightforwardly identify whose norms have been violated, to what degree, or against whom a remedy might be sought. Chapter 6 returns to these limits in detail; the recommendations the dissertation develops do not treat contextual integrity as their sole normative foundation.

Mantelero's (2022) argument about collective and group-level harms extends the normative analysis to dimensions that neither the GDPR nor contextual integrity adequately addresses. When a model trained on web-scraped data systematically encodes or amplifies biases that disadvantage identifiable social groups, whether by gender, ethnicity, religion, or other protected characteristics, the harm is aggregate and diffuse rather than individually particularised. No single person can easily demonstrate that their specific rights were violated to a degree that grounds an enforceable claim, even where the collective harm is substantial. Wachter, Mittelstadt and Russell (2021) make a compatible point in the context of algorithmic fairness law: the opacity of AI systems makes it practically impossible for individuals or regulators to identify and challenge specific rights violations after deployment, because the causal chain between training data, model outputs, and individual outcomes is inaccessible. Individual rights architecture sits poorly with harms of this collective and diffuse character, and that mismatch is one of the structural reasons why governance in this space has consistently produced correct diagnoses without producing effective remedies. The significance of this for the governance argument developed in Section 2.7 is that the problem

cannot be solved merely by improving individual enforcement mechanisms; the design of the enforcement architecture itself requires reconsideration.

2.5 Power, consent and industry self-regulation

The ethical dimension of LLM training on web-scraped personal data operates at a level that legal analysis alone cannot fully reach, and it is worth being clear about the nature of that gap rather than treating ethics as a residual category for cases where law runs out. From a deontological perspective, processing personal data without meaningful consent is a violation of individual autonomy regardless of its legal status. The moral claim people have over information about themselves derives from their standing as rational agents capable of directing their own lives and controlling how they are represented and known to others; it does not derive from what data protection statutes happen to provide. A governance analysis that treats legal compliance as the normative ceiling rather than the floor will systematically under-protect the interests at stake.

Bender et al. (2021) argued, in what has become one of the most widely cited analyses of large language model development, that the costs of building these systems, including privacy harms, substantial environmental impacts, and the encoding of social biases, are systematically externalised onto communities with the least capacity to resist or challenge them. The empirical evidence on exactly who bears those costs is still developing, and it is not always possible to isolate training data extraction as the specific causal mechanism rather than model design and deployment decisions more broadly. The structural power asymmetry the argument identifies is, however, robust: the people whose data is extracted do not consent, are not directly compensated, and have no practical mechanism for challenging the use of their information or its consequences. The benefits flow to developers and, through those developers' commercial products, to users, while the costs are distributed across a largely anonymous population.

Zuboff (2019) frames this dynamic within a broader account of surveillance capitalism, arguing that the extraction of human behavioural and expressive data without consent for commercial purposes represents a form of dispossession: the raw material of human experience and self-expression is appropriated to generate commercial value from which those who produced it are structurally excluded. Hong et al. (2025) provide empirical grounding for this argument at the level of specific data collection practices, finding that

existing pipelines routinely disregard consent or restriction signals expressed through mechanisms such as robots.txt exclusion protocols. The gap between published ethical commitments and actual data collection practices, which this finding illustrates, forms the empirical basis for the conclusion that voluntary commitments do not reliably survive contact with commercial incentives.

Mittelstadt (2019) provides the sharpest theoretical account of why this pattern recurs. He argues that the proliferation of high-level ethical principles across AI governance frameworks, including principles of fairness, accountability, transparency, and privacy appearing in industry codes, government guidelines, and international statements, produces what he terms "ethics-washing": a form of normative legitimisation that allows harmful practices to continue precisely because it creates the appearance of responsible governance without providing the institutional mechanisms necessary to enforce it. The distinction between a principle and an enforceable obligation is not merely technical; it is the difference between a governance framework that changes behaviour and one that provides cover for behaviour that has not changed. Mittelstadt's argument has direct implications for the governance analysis in Chapters 6 and 7: improving the content of ethical principles or the specificity of regulatory guidance will not, on its own, close the gap between declared standards and actual practice. What matters is the institutional design of enforcement, not just the normative content of the rules.

2.6 The adequacy of current governance responses

Regulatory activity has intensified considerably since 2023, and the genuine effort that represents is worth acknowledging before examining the limitations. EDPB Opinion 28/2024 (EDPB, 2024b) is the most comprehensive European statement on data protection in AI model development to date, confirming that the GDPR applies in full to AI training and identifying accountability, documentation, and risk-proportionate safeguarding as operative principles. The ChatGPT Taskforce report (EDPB, 2024a) shows those principles translating into case-specific scrutiny of a major deployment, with findings more practically grounded than abstract guidance alone. The European data protection regulators understand the technical dimensions of the problem and are willing to engage developers on the specifics.

The persistent limitation of both documents (EDPB, 2024a, 2024b) is that they simultaneously acknowledge the practical difficulties of compliance and insist on full

application of the relevant obligations, without resolving the gap between those two positions. That is not a criticism of the EDPB's legal analysis, which is largely correct: the obligations do apply, and the fact that compliance is difficult does not create a legal exception. Guidance that identifies a compliance gap and provides no pathway through it does not generate compliance, however; it generates legal uncertainty and, as Mittelstadt (2019) would predict, provides cover for developers who can claim ongoing engagement with the regulatory framework while their practices remain unchanged.

Technical mitigations are consistently proposed alongside legal regulation as part of an integrated response. Differential privacy, data deduplication, improved anonymisation techniques, opt-out mechanisms, and model unlearning or erasure systems each have genuine merit as partial measures. Their practical limits are, however, significant. Carlini et al. (2023) showed that differential privacy applied at the noise level required to prevent memorisation-based disclosure imposes performance penalties that frontier model developers do not accept in practice: there is an acknowledged trade-off between privacy protection and model capability, and commercial incentives consistently resolve that trade-off against privacy. Kuziemski and Misuraca (2020) argue, in their analysis of AI governance in public-sector contexts, that technical measures generate meaningful protection only when embedded within institutional accountability frameworks that provide external verification and enforcement. Deployed as voluntary industry commitments in the absence of such frameworks, they treat the visible symptoms of a structural problem rather than its underlying cause.

The EU AI Act adds a further layer of complexity. By creating a parallel regulatory framework for high-risk AI systems with its own transparency, conformity assessment, and accountability requirements, it potentially addresses dimensions of the LLM governance problem that the GDPR's personal data focus does not reach. Chapter 6 develops the coordination question in detail.

2.7 Synthesis: why governance keeps failing

The consistent finding across the five themes examined in this review supports a conclusion that none of the individual sources quite articulates directly, and which this section develops as the central analytical contribution of the literature review. The problem with AI governance in this area is not primarily informational. Regulators understand the nature of the problem with considerable precision. The EDPB, CNIL, and ICO have each described the

compliance failures in detail, and the academic literature has provided the empirical and theoretical foundations for that description. The problem is architectural: the enforcement mechanisms available under both the GDPR and the AI Act are designed around individual rights triggers, while the harms produced by LLM training at scale are structurally misaligned with individual rights architecture.

GDPR enforcement ultimately depends on an individual data subject able to identify a rights violation, demonstrate standing, and pursue a remedy. The harms produced by LLM training are not primarily individual: they are collective, diffuse, and systemic. Privacy violations are distributed across populations rather than concentrated in identifiable individuals, and the opacity of training pipelines means most affected people cannot demonstrate the specific harm to their specific data in the evidential terms regulatory proceedings require. Chapter 6 develops the analytical work; the literature review point is that the architecture is the gap.

Mantelero (2022) and Wachter, Mittelstadt and Russell (2021) both point toward this conclusion from different directions, the first from collective rights theory and the second from algorithmic fairness law. Neither, however, draws the governance design implication fully in terms of what enforcement architecture reform would require, and Chapter 6 develops the argument the literature review can only sketch.

In practice, this means governance mechanisms that operate at the systemic rather than individual level: representative enforcement that does not require identifying a single complainant, collective redress for aggregate harms, and proper coordination between data protection authorities and the AI Office. Chapter 7 develops what one such proposal could look like at the level of institutional and technical specificity. The point at the literature-review level is that supplementing the individual-rights framework rather than replacing it is what the literature collectively points toward.

One final observation is methodological. The gap between regulatory knowledge and regulatory effectiveness identified above is reproduced, in a smaller register, within the academic literature itself. Legal scholars work through the GDPR articles without sustained engagement with the empirical evidence; empirical privacy researchers document memorisation and dataset problems without fully drawing out the legal consequences; ethics scholars identify the normative failures without engaging the institutional enforcement mechanisms that would give those findings practical effect. The dissertation treats those three

domains as parts of a single analysis. The methodology chapter develops the integrative approach in detail and acknowledges its risks.

Chapter 3. Methodology

The choice of methodology is not incidental to a dissertation of this kind. The questions the project asks, whether the GDPR's obligations can be met at the scale at which large language models are now built, what protective values are left un-honoured if they cannot, and what governance would have to do to deliver those values, are not questions any single discipline answers on its own. They are doctrinal where the law constrains what can be required of developers, empirical where the answer depends on what those developers actually do, and normative where the law's reach gives out and what privacy protects has to be argued from somewhere else. A methodology adequate to these questions has to bring the three together rather than treat them in series, and the rest of this chapter explains how that has been attempted and what the attempt costs.

3.1 Aim and objectives

The aim of the dissertation is to critically analyse the legal and ethical implications of training large language models on web-scraped personal data, with particular attention to the relationship between the GDPR's requirements and the operational logic of frontier-scale AI development, and to develop institutionally feasible recommendations for AI developers and policymakers.

Five objectives operationalise this aim. They move in sequence from establishing the empirical reality of the problem, through doctrinal legal and normative analysis, to governance recommendations. The first is to establish the empirical reality of personal data presence and memorisation risk in web-scraped LLM training datasets, drawing on dataset audit literature and quantitative memorisation research. The second is to analyse the incoherence of GDPR compliance at AI scale, with particular attention to the legitimate interests balancing test under Article 6(1)(f) and the transparency obligations under Articles 13 and 14. The third is to assess the fundamental rights implications of mass data extraction under Articles 7 and 8 of the EU Charter, including both individual and collective dimensions of harm. The fourth is to evaluate the adequacy of emerging governance frameworks, including the interaction between the GDPR and the EU AI Act, and to identify where the two instruments create enforcement gaps rather than coherent coverage. The fifth is to

develop legally grounded, institutionally feasible recommendations that address the root causes of non-compliance rather than its symptoms.

3.2 Research questions

Three research questions structure the analysis, and they are not independent of one another. The first establishes the legal diagnosis. The second asks whether legal compliance is even the right standard to apply, given the character of the harms involved. The third translates both findings into a governance question.

RQ1 asks whether GDPR compliance is practically unavailable to large-scale LLM development, in the sense that the legitimate interests balancing test under Article 6(1)(f) and the transparency obligations under Articles 13 and 14 presuppose processing conditions that web scraping at scale makes impossible to satisfy. RQ2 asks whether the individual-rights architecture of the GDPR produces a misalignment with the collective and systemic character of the harms LLM training generates, and what this means for whether legal compliance is even the right standard. RQ3 asks what governance reforms, feasible within the existing GDPR and AI Act framework, could address the root causes of the compliance problem and the power asymmetry between AI developers and the people whose data they use.

RQ1 is addressed primarily in Chapter 5. RQ2 runs through Chapter 6. RQ3 connects Chapter 6 and Chapter 7. The three questions correspond directly to objectives O2 through O5. The empirical findings reviewed in Section 2.2 of the literature review supply the factual basis for RQ1, the analysis of the GDPR's structural limitations in Section 2.3 develops the legal diagnosis RQ1 requires, Sections 2.4 and 2.5 provide the normative context for RQ2, and Section 2.6 of the literature review with Objectives O4 and O5 together provide the analytical foundation for RQ3.

3.3 Approach: a desk-based, integrative methodology

The methodology is qualitative and desk-based. It integrates doctrinal legal analysis with interdisciplinary normative inquiry. There is no primary data collection involving human participants. The study draws on three categories of source. The first is regulatory and legal instruments: the GDPR, the EU Charter, EDPB Opinions, decisions of national data protection authorities, and CJEU jurisprudence. The second is empirical academic research

on web-scraped datasets, LLM memorisation, and privacy risk propagation. The third is interdisciplinary scholarship in AI ethics, data protection law, and technology governance.

What makes this methodology integrative rather than additive is that the three bodies of source material are treated as parts of a single analysis rather than as separable domains. Doctrinal legal analysis establishes what the relevant obligations are. Empirical research assesses whether compliance is practically achievable. Normative analysis evaluates whether legal compliance, even if achievable, is sufficient to protect the values the legal framework is designed to advance. None of these can be answered alone. Legal obligations cannot be meaningfully assessed without understanding the empirical context in which they operate, empirical findings cannot be translated into governance recommendations without normative criteria of evaluation, and normative claims about what regulation should do are empty without a careful account of what the law currently requires. The dissertation's contribution lies precisely in joining these literatures, on the premise that treating them separately is what has produced the gap the project addresses.

Integrative work has its own characteristic risks, and naming them is more useful than pretending they do not apply. The first is overreach: a project that draws on doctrinal legal analysis, machine learning research, and political philosophy of privacy is unavoidably operating in fields where deeper specialist work exists, and it can substitute superficial gestures toward each for sustained engagement with any. The second is the related risk of relying on secondary syntheses. It is tempting, when working across literatures, to read each through the lens of an interdisciplinary review article rather than to engage the primary sources, and the resulting analysis is then derivative in ways that are not always visible to the reader. The third is that integrative claims can become unfalsifiable, because the synthesis itself sets the terms on which evidence is read, and a sufficiently flexible synthesis can absorb almost any finding without strain. The mitigations the dissertation has tried to apply are correspondingly modest. Each chapter is anchored in primary literatures of its own field rather than in secondary integrative reviews. Doctrinal claims are made with reference to the regulatory instruments and the case law directly. Empirical claims are sourced from peer-reviewed technical research read on its own terms. Normative claims are made with explicit reference to the philosophical positions they draw on (Nissenbaum's contextual integrity, Mantelero's collective rights argument) rather than as general gestures toward "the literature on AI ethics". Where integrative claims connect these, the connection is shown rather than asserted, and where the synthesis goes further than any single literature licenses, the chapter

says so. None of this is a complete answer to the methodological objection. It is the best a single-student undergraduate project can do within the constraints of a desk-based study.

A connected risk that an integrative methodology raises in particular form is selection bias in the source base. A project that draws on technical, doctrinal, and normative literatures has, in principle, a much wider field to choose from than a single-discipline project, and the choices made in narrowing that field deserve to be on the record rather than left implicit. The principles applied here were three. Technical sources were drawn from the peer-reviewed memorisation and dataset-audit literature on which the empirical chapters depend, with Carlini et al. (2021, 2023), Nasr et al. (2023), Dodge et al. (2021) and Hong et al. (2025) forming the core. Where alternative findings exist that contradict the broad pattern those sources establish, the dissertation has not been able to locate them, and Section 3.6 notes Hong et al.'s status as an unreviewed preprint. Doctrinal sources were drawn from the regulatory instruments and case law directly, supplemented by the EDPB, CNIL and ICO guidance that frames their application to LLM training. The selection here is comprehensive rather than illustrative: every major European regulator's position over the period covered has been engaged. Normative sources were chosen for their direct relevance to the protective values the analysis develops, with Nissenbaum, Mantelero, Solove and Mittelstadt each carrying a specific argumentative load that Chapter 6 makes visible. Critics writing from positions less sympathetic to the dissertation's thesis, including those defending the procedural reading of Article 5 or the public-private distinction in Charter case law, have been engaged where their arguments bear on the analysis rather than treated as a separate body of "opposing literature". The remaining risk of selection bias is one the dissertation cannot fully discharge, and the appropriate response is the one taken throughout: state the position being defended explicitly, name the places where it depends on contested readings, and let the reader weigh the analysis against the alternatives the chapters flag.

3.4 Why not other approaches

Three other approaches were considered for the project before the desk-based integrative one was settled on, and the case for the chosen approach is sharper if the rejected ones are taken seriously rather than waved at.

The most obvious alternative was a directly empirical one: pick a publicly available web-scraped corpus, run a sample audit for personal data prevalence, and probe a model trained on

it for memorisation. This was rejected for two reasons that are not the same. The empirical ground is already covered. Hong et al. (2025) and Carlini et al. (2021, 2023) provide the quantitative evidence base the dissertation needs, and replicating that work at smaller scale would add little. The gap in the literature is not more empirical data but analysis of what the existing evidence means for the legal and governance frameworks. Even if the gap had pointed the other way, conducting a meaningful dataset audit within a single-student project on an eight-month timeline is not realistic. The corpora are measured in terabytes, and the compute and infrastructure required for serious memorisation testing exceed what an undergraduate dissertation can reasonably command.

Comparative analysis across multiple jurisdictions was a more tempting option. Setting the GDPR alongside US sectoral data protection law, the UK's post-Brexit regime, and Chinese personal information protection law would have placed the European framework in useful comparative perspective. The reason against was structural rather than dispositive. Comparative work of any analytical depth would have meant reducing each jurisdiction to a chapter or less, and the regulatory positions in three of the four jurisdictions were much less developed than the European one over the period the dissertation covers. For a project of this size, depth in the jurisdiction where enforcement decisions, regulatory opinions, and academic commentary are actually accumulating produces sharper analysis than breadth across several.

Practitioner interviews, with data protection officers, regulatory staff, or developer-side counsel, would have added useful context on how the compliance challenges are experienced from inside the institutions involved. The reason against was that the central research questions are doctrinal and normative rather than empirical. Whether GDPR compliance is structurally available to large-scale LLM development is a question answered by analysing legal instruments and regulatory guidance, not by asking practitioners whether they find compliance difficult. Practitioner interviews would have added useful context but would not have addressed the core argument.

3.5 Positionality

The position from which this dissertation is written matters for the analysis, and it is more useful to name it than to pretend to a neutrality I do not have. I am a final-year Computer Science student, not a legal scholar, and I have worked through the doctrinal sources of EU

data protection law from the outside rather than from inside the legal academy. That has affected the analysis in particular ways. The technical chapters, especially the empirical landscape in Chapter 4 and the discussion of memorisation and dataset composition, are written with confidence that comes from familiarity with how machine learning systems are actually built. The legal chapters are written with the care of someone reading primary sources directly and being aware that the doctrinal commitments they take could be qualified by litigators and academics who have spent years on this material. Where I have not been confident in a doctrinal reading, I have said so, and Chapter 7's limitations section returns to which specific commitments are most exposed to subsequent revision.

The working hypothesis I came to the project with, and which the analysis has tested rather than presupposed, is that the regulatory framework is not failing for want of clearer guidance or stronger enforcement will. It is failing because the enforcement architecture is built for a kind of processing relationship that frontier-scale LLM training does not have, and the individual-rights mechanisms cannot reach the collective and diffuse harms the activity actually produces. That hypothesis shaped the framing of the research questions and the direction the governance argument develops in. It would be dishonest to pretend the dissertation arrived at this conclusion neutrally. What it can claim is that the conclusion has been tested against the strongest objections I have been able to find, and that the chapters say where I have been persuaded and where I have not.

3.6 Limitations

The methodology has limits the analysis cannot transcend, and naming them honestly is part of doing the work properly. The most consequential is that the regulatory position in this area is moving faster than a dissertation can keep up with. The European Commission's implementing acts under the AI Act are still in train at the time of writing, the EDPB has guidance on the GDPR/AI Act interface still in preparation, and the Court of Rome's annulment of the Garante fine in March 2026 has not yet been followed by the appellate response that will determine its eventual reach. The analysis works with what is available as of early 2026, and a reader picking the dissertation up six months later will be reading an account that has not seen what came next. The broader architectural argument depends less on the details of any single instrument than on the structural pattern across them, but specific legal conclusions, particularly in Chapter 6's case study sections, are exposed to subsequent regulatory developments in ways the chapter itself flags.

The empirical literature the dissertation rests on has its own constraints worth flagging. The technical work on memorisation and dataset composition (Carlini et al., 2021, 2023; Nasr et al., 2023; Dodge et al., 2021; Hong et al., 2025) is specialised, peer-reviewed in most cases, and consistent across studies in its broad findings. Hong et al. is, at the time of writing, an unreviewed preprint, and the weight given to it reflects the consistency of its findings with the broader empirical record rather than independent confirmation. The studies examine specific corpora and specific model architectures. Generalising their findings to all LLM development pipelines involves some extrapolation, and a marker who pushes hard on this point would not be wrong. What the dissertation can offer in response is that the legal arguments do not depend on quantitative precision in the empirical findings: what matters is that memorisation is a real phenomenon at frontier scale, not what its exact rate is in any particular system.

Two further constraints are worth registering briefly. The geographic focus on the EU and the UK is defensible because the regulatory and academic activity worth analysing has been concentrated there, but a fuller account of global regulatory divergence is beyond what a single dissertation can sustain. The analysis also works exclusively from what is publicly disclosed; where actual practice diverges from public disclosure in ways that have not surfaced through regulatory investigation, the analysis cannot account for it. Neither constraint undermines the structural argument.

Chapter 4. The Empirical Landscape

4.1 Introduction

The literature review established that personal data is present in web-scraped training corpora because that is what the open web contains: people write about themselves online, and any corpus built from the open web at scale will contain personal data as a result. The legal analysis that follows turns on a more detailed picture of the LLM training pipeline as a regulatory object than the literature review had room for. This chapter develops that picture by reconstructing from the existing literature how web-scraped data becomes a deployed model, in enough detail for the next chapter to identify where each compliance obligation breaks down. No new empirical findings are presented; the synthesis is the contribution.

One framing point matters at the outset. The chapter treats the LLM training pipeline as a regulatory object, not just an engineering one, and that framing is itself contested. The Hamburg DPA (2024) and the EDPB (2024b) disagree at a fundamental level about whether the trained model contains personal data at all, and that disagreement is not a peripheral doctrinal squabble. It is the question that determines whether GDPR rights apply to deployed LLMs. Section 4.5 returns to it as the central unresolved issue the empirical landscape produces for the legal framework.

4.2 The training pipeline as a regulatory object

Regulators tend to refer to "training data" as if it were a discrete, inspectable thing: a defined corpus that a developer holds, processes, and could in principle expose to a data subject access request. The engineering reality is less tidy than that. The legal character of the data arguably changes at each stage of a multi-stage pipeline, and identifying where a given GDPR obligation is meant to bite within that pipeline is not straightforward. Several of the doctrinal difficulties developed in Chapter 5 turn on precisely this point.

The first stage is acquisition. For models trained on the open web, this almost always begins with Common Crawl, a non-profit web archive that has been crawling the public web since 2008 and now hosts petabytes of raw HTML across hundreds of billions of pages (Common Crawl Foundation, 2024). Common Crawl itself does not curate or filter. It functions closer to a public-record snapshot of the indexable web. Derived datasets are then constructed from

this raw substrate by applying language detection, deduplication, quality filtering, and content blocklists. C4, used to train Google's T5 and Switch Transformer models, was constructed in this way (Raffel et al., 2020). The Pile (Gao et al., 2021) supplements Common Crawl with twenty-two further sources including PubMed Central, ArXiv, Stack Exchange, and the Books3 corpus. RefinedWeb (Penedo et al., 2023), Dolma (Soldaini et al., 2024), and FineWeb (Penedo et al., 2024) represent more recent attempts to build better-curated derivatives. Individual labs maintain proprietary variants, but the underlying source material is largely shared across the field.

Preprocessing comes next. Documents are deduplicated, tokenised into sub-word units, and shuffled into training shards. The original text, after this, exists only as a sequence of integer indices into the model's vocabulary. The tokenised corpus is then fed through the optimiser, which adjusts billions of model parameters to reduce next-token prediction error across the corpus. The original documents are not present in the model after this stage. What remains is a fixed-size weight matrix.

Deployment is the last stage. The trained weights are loaded into an inference server and exposed to users through an API or chat interface. User inputs are tokenised in the same way training inputs were, and outputs are generated probabilistically by sampling from the model's predicted next-token distribution.

Two features of this pipeline matter for the analysis that follows. The relationship between an individual document and the deployed model is many-to-one and lossy. Hundreds of billions of training documents are compressed into a model whose parameter count, although large, is several orders of magnitude smaller than the training corpus. The intuitive picture is that personal data ought to dissolve in this process, and that intuition has done a substantial amount of work in industry communications. The empirical picture, as Section 4.4 develops, is that it sometimes does not. The other feature is that the boundary between "the dataset" and "the model" is in part a regulatory choice rather than a technical fact. The Hamburg DPA's argument that LLMs do not store personal data depends on locating the personal data exclusively in the first stage and treating the second and third stages as transformations that strip the data of its identifying character. Whether that location is sustainable is the question the chapter has to confront.

4.3 What web-scraped corpora actually contain

The most thorough public investigation of a major web-scraped corpus remains Dodge et al.'s (2021) audit of C4. The study itself was not framed in regulatory terms, but its findings bear directly on the regulatory analysis. Dodge et al. found that C4's apparent neutrality conceals a dataset whose composition is heavily shaped by pipeline choices rather than by deliberate selection. The most common top-level domains in C4 included patents.google.com and US military websites, which reflects the filtering pipeline's preference for long, formally structured, English-language text rather than any considered editorial judgement. The blocklist used to filter "bad" content disproportionately removed material written in or about minoritised dialects and identities, including non-pornographic discussion of LGBTQ+ topics. Substantial volumes of machine-generated text and benchmark evaluation data were also identified. The latter creates well-known evaluation contamination problems, but it also indicates that the corpus contains material whose provenance the dataset's users could not have reliably identified. The dataset is, in the language of Dodge et al. (2021), a sociotechnical artefact whose character reflects the filtering pipeline at least as much as the underlying web.

Birhane and Prabhu's (2021) audit of the LAION-400M dataset, although focused on a multimodal rather than a text-only corpus, established a related empirical claim with sharper implications. They documented that LAION contained explicit non-consensual sexual imagery, racist and misogynistic content, and identifiable images of named individuals. LAION-5B, the successor dataset used to train Stable Diffusion, was subsequently found by Stanford researchers to contain confirmed child sexual abuse material and was withdrawn (Thiel, 2023). The LAION case is sometimes treated as separate from the LLM debate because it concerns image data, but the conceptual point is the same. A dataset that aggregates the open web at scale will reflect the open web, including its illegal and abusive content, regardless of the developers' intentions or stated filtering policies.

The studies discussed in Chapter 2 add a further dimension to this picture. Hong et al. (2025) audited a prominent web-scraped corpus and identified persistent volumes of personally identifiable information including medical records, financial information, and data concerning minors, even after sanitisation had already been applied. The EDPS (2023) reached the same conclusion at a higher level of abstraction, framing privacy exposure at the collection stage as architecturally inevitable for state-of-the-art systems rather than contingent on any particular design choice.

The natural objection to this picture is that improved filtering will, over time, reduce these problems. On this view, the residual presence of personal data is a contamination issue rather than a structural feature. Major developers have adopted this framing in their public communications, and some commentators in the policy literature have followed them. The empirical record gives several reasons for treating the framing sceptically. The trajectory of dataset construction is towards larger and less aggressively filtered corpora rather than smaller and more curated ones. The authors of RefinedWeb explicitly argued that aggressive filtering is counterproductive at scale because it removes useful linguistic variation along with undesirable content (Penedo et al., 2023), and frontier model performance has tracked dataset size more closely than dataset quality in ways that create a structural incentive against selective curation. The kinds of personal data that are most legally sensitive are also precisely the kinds the open web contains in large volumes. People discuss their medical conditions, sexuality, political opinions, and religious beliefs online. A filter aggressive enough to remove all such content would remove a substantial portion of the corpus and degrade the resulting model. The filters that are applied operate on textual surface features rather than on the legal status of the data. A document discussing someone's medical condition is not flagged by a quality filter, and a piece of leaked correspondence is not flagged by a blocklist. The filtering pipeline is largely orthogonal to the data protection categories that GDPR uses to allocate legal protections, and that orthogonality is what makes the structural framing more defensible than the contamination one. Residual problems can in principle be engineered away. Structural problems can only be addressed by changing the activity itself.

4.4 From corpus to model: how privacy risk travels

The transition from training corpus to trained model is where the regulatory analysis becomes most contested. Two empirical claims have to be established carefully before the legal stakes become clear. The first is that LLMs memorise portions of their training data. The second is that this memorisation is exploitable in deployed production systems rather than only in research models. Both claims are now well documented in the empirical literature, but the legal significance of each is different and has not been fully absorbed into regulatory thinking.

The memorisation claim was first established at scale by Carlini et al. (2021), whose study of GPT-2 demonstrated that training data could be extracted verbatim from a deployed model under specific prompting conditions, including content containing personally identifiable

information. Carlini et al. (2023) extended this work into a quantitative study across a range of model sizes and training regimes. They showed that memorisation scales with three factors: model parameter count, frequency of duplication in the training data, and the length of the prompt context provided to the model. Their headline finding, that the largest and most capable models memorise the most, has been independently replicated and is now accepted across the field. Debates remain about whether the rates observed under research conditions translate into practically meaningful exposure in deployed systems, but the underlying phenomenon is no longer in dispute.

The exploitability claim was the contribution of Nasr et al. (2023). Their "divergence attack" against ChatGPT caused gpt-3.5-turbo to abandon its chat-style behaviour and emit training data at a rate the authors estimated to be 150 times higher than under normal use. The extracted material included email addresses, phone numbers, full names, and substantial verbatim passages from copyrighted texts. Several features of this finding deserve attention. The attack worked despite reinforcement learning from human feedback, which had been assumed to suppress memorisation as a side effect of alignment. The authors disclosed the vulnerability to OpenAI and waited 90 days before publication, which the responsible disclosure community treats as confirmation that the issue is genuine and exploitable rather than artefactual. The specific divergence prompt that triggered the attack has been patched, but the authors note that the underlying vulnerability, the model retaining extractable training data, has not been fixed, because it cannot be fixed without retraining the model from scratch on differently constructed data.

Taken together, the findings of Carlini et al. (2021, 2023) and Nasr et al. (2023) support a claim the regulatory analysis has not fully absorbed. Memorisation is not a rare malfunction. It is a predictable consequence of the training objective itself. Models are trained to minimise prediction error on the training distribution, and for sequences that appear repeatedly or that are unusual enough to be informative, the optimiser pushes the model toward reproducing them. Personal data on the open web is often duplicated and often unusual in the relevant sense. A specific phone number is, by construction, a low-probability sequence, and is therefore disproportionately likely to be memorised. The legal significance is that the harms documented in the extraction literature are not edge cases in poorly engineered systems. They are the foreseeable output of standard training practices applied to standard training corpora. It is worth flagging from a system-design perspective that the trade-off between memorisation risk and model capability is not a privacy-versus-performance afterthought but

a real engineering constraint. Differential privacy at the noise levels required to suppress memorisation-based extraction degrades model performance to a degree that frontier developers have repeatedly judged commercially unacceptable, and the trade-off has not yet yielded to better technique. The legal arguments that follow do not assume this constraint is permanent, but they do take seriously that engineering against memorisation is materially harder than engineering for it, and that this asymmetry shapes which compliance routes are realistic.

The standard industry response to this picture is to invoke anonymisation as a route out of GDPR's scope. The argument is that whatever the input data was, the resulting model has so transformed the data, dispersing it across billions of parameters in numerical form, that the resulting object is anonymous and falls outside the regulation. The empirical findings just summarised place this argument under significant pressure. If the model can be made to emit verbatim training data containing identifiable individuals, the claim that the data has been irreversibly anonymised is, on its face, difficult to sustain. The more defensible version of the industry argument is that the memorisation rate is sufficiently low to count as a residual error rather than a defining feature, but that defence depends on a quantitative threshold the literature has not converged on. Carlini et al. (2023) report extraction rates that vary substantially with model and prompt. Nasr et al. (2023) show that aligned production systems can be made to leak data at much higher rates under adversarial conditions. Whether the threshold for anonymity under GDPR sits below or above these rates is the contested question Section 4.5 develops, and which regulators are now actively litigating.

4.5 The deployed-model question and the Hamburg-EDPB disagreement

The literature review raised a question the chapter is now in a position to develop more fully: whether deployed model weights containing memorised personal data constitute "processing" within the meaning of Article 4(2) GDPR. That question has acquired sharp practical importance because two prominent regulatory authorities have taken opposing public positions on it, and the disagreement reveals a deeper conceptual tension in how the GDPR's definitions apply to machine learning systems.

The Hamburg DPA's discussion paper of July 2024 advances three theses. The mere storage of an LLM does not, in its analysis, constitute processing within the meaning of Article 4(2) GDPR, because no personal data is stored in the model. Where personal data is processed by

an "LLM-supported AI system", for example through user inputs and outputs, that processing must comply with GDPR, but the obligation attaches to the system using the model rather than to the model itself. Data subject rights under Articles 15 to 22 cannot apply to the model itself, because there is no personal data within the model on which they could operate (Hamburg DPA, 2024). The argument's technical foundation is that LLMs store correlations between tokens in the form of high-dimensional weight vectors. These vectors cannot, in the Hamburg DPA's view, be linked back to specific identifiable individuals in the manner CJEU jurisprudence requires for "information relating to a natural person". The Danish Datatilsynet has reached a similar conclusion (Datatilsynet, 2023), so the Hamburg position is not isolated within the regulatory community.

The European Data Protection Board's Opinion 28/2024, adopted on 17 December 2024 in response to a referral from the Irish Data Protection Commission, takes a markedly different position. The EDPB declined to hold that AI models trained on personal data are categorically anonymous. Instead, it set out a two-pronged test: an AI model can be considered anonymous only where both the likelihood of direct, including probabilistic, extraction of personal data about training-set individuals, and the likelihood of obtaining such data from queries, are insignificant taking into account all means reasonably likely to be used (EDPB, 2024b). The Opinion explicitly grounds this position in the memorisation literature, observing that LLMs can inadvertently memorise and leak pieces of their training data even where they were not designed to do so, and concluding that any AI model from which identifiable data can be extracted or reconstructed cannot be classified as anonymous. The case-by-case framing leaves room for some models to be treated as anonymous, but the threshold has been set in a way that, given the empirical findings developed in Section 4.4, frontier LLMs are unlikely to meet under their current training and deployment regimes.

The disagreement between Hamburg and the EDPB is more than a difference of regulatory style. The two authorities are answering different questions about what an LLM actually is, in legal terms. Hamburg treats it as a question about the artefact: open the model file, look at the weights, and ask whether the numbers stored there contain "information relating to" an identifiable person. On a narrow reading of "contain", they do not. Nowhere in the weights is there a record saying "X has medical condition Y". The EDPB treats it as a question about behaviour: prompt the deployed system, observe what it outputs, and ask whether the model can be made to produce information about identifiable individuals as part of its normal operation. On the empirical record, it can. The two positions diverge because they have

located the data protection question at different points in the pipeline, and the GDPR's text plausibly supports either location.

The IAPP's critique of the Hamburg position (IAPP, 2024) is worth engaging with at this point because it sharpens the analytical issue. The Hamburg argument depends on a narrow technical definition of personal data that is in tension with longstanding CJEU case law treating GDPR's concepts as functional rather than technical. On the functional reading, information can qualify as personal data not only by virtue of its content but also by virtue of its purpose or its result, that is, whether it is used to evaluate or affect identifiable individuals. A corpus of model weights that can, on appropriate prompting, produce information about identifiable individuals fits the functional definition of personal data even if it does not, on inspection, contain anything a human reader would recognise as such. The IAPP critique suggests, with some force, that the Hamburg position works only on a textually narrow reading the CJEU has previously declined to adopt.

The disagreement connects directly to RQ1. If the Hamburg view is correct and trained models do not contain personal data, then most of GDPR's apparatus does not apply to deployed LLMs. Data subject rights drop out, the right to erasure has no purchase, retention obligations attach only to raw corpora rather than to the system trained on them. The lawful-basis question would still bite at the training stage, but the deployed model would be regulatory dark matter: built from personal data, capable of producing personal data, and itself outside data protection law. The most consequential operation in the pipeline would carry the lightest regulatory load. A model deployed to millions of people for years would be less regulated than the corpus it was trained on. That is an inversion of the GDPR's logic, and the EDPB Opinion can be read as an attempt to prevent it by holding the line at the functional reading, though the case-by-case framing leaves room for that line to erode in practice.

The disagreement does not need to be settled here. What the doctrinal analysis in the next chapter does need is a working position, and the working position is the EDPB's. The empirical record of memorisation and extraction provides strong reasons to read the model functionally rather than as an inert numerical artefact, and the Hamburg view, even on its most charitable reading, makes peace with an outcome that empties most of the GDPR's apparatus of work in the deployed-model context. The Hamburg position has been adopted by a serious supervisory authority and points to a real difficulty the GDPR has in conceptualising machine learning systems. The dispute remains live. The broader argument

that follows does not depend on which side ultimately prevails, and where the analysis does depend on the EDPB reading, the chapter will say so.

4.6 Empirical premises for the doctrinal chapter

The picture this chapter has built has several connected features the next chapter draws on. Web-scraped corpora at frontier scale contain substantial personal data including special-category data, as a structural consequence of how the open web operates rather than a contamination problem amenable to better engineering. Training does not reliably break the link between input and model: memorisation scales with model size and is exploitable in deployed production systems. The legal status of the trained model is contested, with the empirical record favouring the EDPB's functional reading over Hamburg's narrow technical one. The pipeline's structure means GDPR obligations attaching to processing operations cannot be mapped to discrete points in the developer's workflow without significant strain.

These features support the central thesis without establishing it. The doctrinal work of the next chapter is required to show how each obligation breaks down on its own terms. What the empirical landscape establishes is that the conditions any doctrinal argument now has to be made under are not the conditions GDPR's drafters appear to have anticipated. That gap is the substantive subject of what follows.

Chapter 5. GDPR Analysis

5.1 Introduction

The literature review identified a set of compliance difficulties that GDPR poses for large-scale LLM development and surveyed the academic and regulatory responses to them. The picture that emerged was one of an obligation framework under significant strain. Lawful basis, transparency, and accountability mechanisms designed for bounded data-processing relationships are being asked to govern a form of processing whose data subjects are unknown, whose purposes are open-ended, and whose technical character does not map cleanly onto the regulation's implicit data model. The literature review treated these as connected difficulties but did not have space to develop the doctrinal analysis that would show how each individual obligation actually fails. This chapter undertakes that analysis.

The chapter focuses on the obligations the literature review had room only to gesture at: Article 9 on special-category data, Articles 15 to 22 on data subject rights, and Article 5 read with Article 25 on processing principles and design.

The chapter does not revisit the lawful-basis and transparency analyses developed in Section 2.3, except briefly in Section 5.2 to set up the chapter's own contribution. The empirical premises the chapter relies on were established in Chapter 4, including the Hamburg-EDPB disagreement that determines whether the deployed model is itself within scope of GDPR's obligations. Chapter 5 reasons throughout from the EDPB's functional-definition reading, on which trained models from which identifiable data can be extracted are not anonymous and remain subject to the regulation, while noting where the Hamburg view would alter the analysis.

5.2 What the literature review established

Section 2.3 set out the analytical core of the lawful-basis and transparency problems: the Article 6(1)(f) balancing test cannot be conducted with substantive content where data subjects are unknown, and Articles 13 and 14 cannot be met at scale. Kamarinou, Millard and Singh (2017) located the difficulty in a deeper structural friction between purpose limitation and data minimisation and the data-maximisation logic of modern AI systems, and Solove

(2021) framed the transparency failure as the obsolescence of notice-and-consent frameworks built for bounded processing relationships.

What the literature review did not develop is how these difficulties intersect with the obligations the chapter now takes up: Article 9 on special-category data, Articles 15 to 22 on data subject rights, and Article 5 read with Article 25 on processing principles and design. Each is both more important and more underdeveloped in the existing literature than the lawful-basis and transparency obligations that have attracted the most attention. Working through them is what allows the chapter to move from "several GDPR obligations are difficult to comply with" to the stronger structural claim the dissertation defends.

5.3 Article 9 and the limits of "manifestly made public"

Article 9(1) GDPR prohibits the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health data, or data concerning a natural person's sex life or sexual orientation. The prohibition is subject to the exceptions in Article 9(2), of which the only one with any plausible application to LLM training on web-scraped data is Article 9(2)(e): where "processing relates to personal data which are manifestly made public by the data subject". The doctrinal question is whether this exemption can carry the weight that the empirical conditions described in Chapter 4 would require it to.

The empirical setup is straightforward. Special-category data is on the open web in substantial volumes. People discuss their medical conditions in support forums, disclose their religious beliefs and political opinions on social platforms, write about their sexuality in personal blogs, and document health data in patient communities. Hong et al. (2025) identified medical records in a sanitised web-scale corpus, and the EDPS (2023) framed special-category exposure at the collection stage as architecturally inevitable for state-of-the-art systems. Filtering does not solve the problem because, as Section 4.3 developed, the kinds of data that are most legally sensitive are precisely the kinds the open web contains because people discuss the relevant aspects of their lives online. Any model trained on a sufficiently large web-scraped corpus will, with near certainty, have processed special-category data within the meaning of Article 9(1).

The question, then, is whether Article 9(2)(e) provides a route through. The CJEU's interpretation in GC and Others v CNIL (Case C-136/17, 2019) is the most relevant authority.

The Court held that the "manifestly made public" exemption requires the data subject to have, with full awareness, explicitly made the data public, and that the exemption is to be read strictly because Article 9 is itself an exception to a general protective regime. The Court rejected the argument that data is manifestly made public just because it is accessible on the open internet. The data subject must have positively intended publication of the special-category data, with awareness of what they were publishing. This is a much higher threshold than ordinary public availability.

The implications for LLM training are significant. A user who posts on a health forum about their medical condition has indeed made that information public, but the question Article 9(2)(e) asks is whether they made it "manifestly" public in the strict sense the CJEU requires, with awareness that "made public" entails being available for any subsequent processing including processing that did not exist at the time of disclosure. The answer is, on the strict reading, no. The user disclosed the data within a particular communicative context with particular reasonable expectations, none of which included its incorporation into the training data of a foundation model that did not exist when the disclosure was made. The contextual integrity literature (Nissenbaum, 2019) has long argued that public availability and unrestricted availability for any purpose are not the same thing, and the CJEU's strict reading of Article 9(2)(e) appears to incorporate that distinction at the doctrinal level.

There is a more developer-friendly reading on which Article 9(2)(e) is satisfied whenever the data subject has chosen to publish the data in a publicly accessible location. This reading would treat the strict CJEU jurisprudence as applicable only to specific contexts, such as the search engine context in GC and Others, and would leave room for a broader application in the LLM context where the volumes involved make individual assessment impossible. The question is whether the fragmentation of doctrine such a reading would require is legitimate or whether it amounts to reading the exemption to mean something the CJEU has explicitly said it does not mean. The strict reading is the doctrinally defensible one. The CJEU's reasoning in GC and Others does not turn on features specific to search engines but on the structure of the exemption itself, and applying a more permissive reading at LLM scale would invert the regulation's logic by making the exception swallow the rule whenever processing occurs at sufficient volume.

The structural consequence is that Article 9 sits in a position where the empirical reality of the activity and the doctrinal reality of the regulation are simply incompatible. Frontier-scale

web-scraped corpora unavoidably contain large volumes of special-category data. Article 9(2)(e) on its strict reading cannot legitimate the processing of that data, and no other Article 9(2) exemption applies. The literature review's diagnosis of the legitimate-interests problem found a similar incompatibility at the Article 6 level but identified it as a balancing-test failure that procedural mechanisms could not cure. Article 9 makes the same kind of incompatibility visible at a more fundamental layer. There is no balancing test under Article 9, only a binary prohibition with narrow exceptions, and the exceptions do not fit the activity.

5.4 Data subject rights in a memorisation regime

The data subject rights established in Articles 15 to 22 are the GDPR's principal mechanism for translating its protective principles into individual remedies. On the EDPB's reading developed in Chapter 4, these rights apply to deployed LLMs because such models are not anonymous in the sense Article 4(1) requires. The question this section examines is what the application of those rights actually looks like in practice, and whether the rights as drafted can be discharged at all in the technical conditions LLM deployment creates.

Article 15 is the access right and it is the obvious one to start with, because it is the mechanism by which a data subject can find out what is happening to their data in the first place. It gives the data subject the right to confirmation that their personal data is being processed and, if so, a copy of that data along with information about purposes and recipients. The mechanics of access against a deployed LLM are not encouraging. The model does not maintain an index of the individuals whose contributions shaped it, and there is no procedure that takes a name and returns whatever the model has memorised about that person. A controller could try to elicit such information by prompting the model in particular ways, but a generative system queried for facts about a named individual will sometimes produce accurate memorised content, sometimes plausible fabrication, and often a mixture. The output is not a reliable answer to the access question, and the act of producing it is itself further processing of personal data. The right is in force on the EDPB's reading. The conditions under which it could actually be discharged are not.

Rectification under Article 16 is sharper. A data subject is entitled to have inaccurate personal data about them corrected without undue delay, and the obligation makes immediate sense for systems that hold records: find the wrong record, change the wrong field, save the new value. Trained models are not such systems. There is no record to find, and the

activation patterns that produce a misstatement about an identifiable individual are diffused across billions of parameters. The interventions actually available are managing what the model says rather than correcting what the model is. The system can be fine-tuned to produce different outputs in response to specific prompts. An output filter can be added downstream. Neither operation rectifies anything in the sense Article 16 contemplates, because neither touches the underlying weights. The Italian Garante's 2023 order against OpenAI (Garante, 2023) made the point concrete in practice. The regulator required OpenAI to provide mechanisms for users to correct inaccurate information, and what eventually emerged was a complaint-handling pipeline rather than rectification in any technically meaningful sense.

Erasure under Article 17 is the sharpest version of the same problem. The right entitles a data subject to have their personal data erased without undue delay where, among other grounds, the data is no longer necessary for the original purposes or where consent has been withdrawn. To erase a particular individual's contributions from a frontier-scale LLM would mean retraining the model on a corpus that excludes them, which is a multi-month, multi-million-dollar undertaking, and it is not plausible that erasure requests will trigger such retraining at any meaningful frequency. The literature on machine unlearning (Bourtoule et al., 2021; Nguyen et al., 2022) has explored more efficient approximations, but the techniques are not yet operational at frontier scale and the guarantees they offer are statistical rather than verifiable. Article 17 contemplates erasure that the data subject can rely on. What current engineering can offer is a statistical assurance that the model is less likely to produce the relevant content. The gap between those two is not a question of refinement.

Article 20 fails for a different reason. The portability right entitles a data subject to receive, in machine-readable form, the personal data they have provided to a controller. The phrase that does the work is "provided to". Web-scraped training corpora contain data the data subject did not provide to the controller at all. They posted it elsewhere, on a forum, a blog, a social platform, in some other context, and a scraper picked it up. Article 20's scheme presupposes a bilateral provision relationship that web-scraped corpora do not instantiate, and the right does not formally apply where that relationship is missing. The deeper point is not that Article 20 happens to fail in this case but that its underlying picture of how data ends up with a controller does not match how training corpora are built.

The pattern across these rights is consistent. Each runs into a different operational impossibility, and the impossibilities are not coincidental. The rights were designed for

systems that hold individual records in addressable form, where access can be granted by retrieving the record, rectification by editing it, erasure by deleting it, and portability by exporting it. LLMs are not such systems. They distribute the influence of any given record across billions of parameters in a way that breaks the operational presupposition of each right. The point is not that individual workarounds cannot be engineered. Some can be, partially. The point is that the rights, taken as a coherent protective scheme, presuppose a data architecture that LLMs do not have, and the scheme degrades when that architecture is absent.

5.5 Article 5 principles and the structural-incompatibility argument

Article 5 establishes the GDPR's processing principles: lawfulness, fairness and transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability. Article 25 then makes data protection by design and by default a binding obligation on controllers, which means the principles in Article 5 have to be built into the design of any processing operation rather than bolted on afterward. Together, these two articles are where the GDPR's protective philosophy actually lives, and they are where the structural-incompatibility argument the chapter has been building toward bites hardest.

The purpose limitation principle in Article 5(1)(b) requires that personal data be collected for specified, explicit, and legitimate purposes and not further processed in a manner incompatible with those purposes. Foundation-model training does not have a specified purpose in the sense the principle contemplates. The entire commercial logic of foundation models is that they are trained once on broad corpora and then deployed for purposes that are not specified at training time and could not be specified, because the model's deployment downstream is itself open-ended. The literature review identified this difficulty in Kamarinou, Millard and Singh (2017), who diagnosed it as embedded in the design of both the law and the technology. The doctrinal formulation is sharper. Article 5(1)(b) cannot be complied with by an activity whose entire economic rationale is purpose-flexibility.

Data minimisation under Article 5(1)(c) requires that personal data be adequate, relevant, and limited to what is necessary in relation to the purposes for which it is processed. Foundation-model training is the inverse of data minimisation. The training procedure works better with more data, and the empirical findings on scaling laws (Hoffmann et al., 2022) confirm that

data volume is one of the key determinants of model capability. A controller asked to demonstrate compliance with Article 5(1)(c) would have to show that the training data included only what was necessary for the model's purposes, but the training procedure does not operate by selecting necessary data. It operates by aggregating as much data as feasible and allowing the optimiser to determine which parts are useful. The "necessary" framing does not fit the activity.

Storage limitation under Article 5(1)(e) requires that personal data be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which it is processed. Training data in modern LLM development is retained indefinitely, both in raw corpus form and embedded in the weights of trained models, with no mechanism for time-bounded deletion of individual contributions. The model's value depends on the persistence of the training that produced it. There is no operational sense in which training data is held only "for as long as necessary" because the necessity is, by construction, the indefinite operating life of the deployed model.

The accuracy principle in Article 5(1)(d) requires that personal data be accurate and, where necessary, kept up to date. LLMs notoriously generate confident-sounding but inaccurate statements about identifiable individuals, sometimes called hallucinations in the technical literature. The Italian Garante's order against OpenAI cited inaccurate outputs about identifiable individuals as one of its grounds (Garante, 2024). The accuracy principle was drafted for systems that hold records about individuals which can be checked and corrected. Generative systems do not hold records in that sense. They generate plausible continuations from learned distributions, and the accuracy of those continuations is a property of the generation process rather than of any underlying record. There is no procedure that would allow a controller to ensure the accuracy of an LLM's outputs about identifiable individuals, because the outputs are not retrievals from a corrected record.

Article 25 brings these difficulties together. Data protection by design and by default requires the controller to implement technical and organisational measures that give effect to the data protection principles, taking into account the state of the art, the cost of implementation, and the nature, scope, context, and purposes of the processing. The literature has tended to treat Article 25 as a procedural obligation, requiring controllers to conduct DPIAs, document design choices, and demonstrate that protection has been considered at the design stage. On a procedural reading, Article 25 can be complied with by sufficiently rigorous documentation.

The deeper reading is that Article 25 requires the principles in Article 5 actually to be honoured at the design level, and where the activity's design is structurally incompatible with the principles, no documentation cures the substantive failure. The structural-incompatibility argument the chapter has been developing reaches Article 25 as its culmination. An activity that cannot honour purpose limitation, data minimisation, storage limitation, or accuracy in any substantively meaningful sense cannot be designed to give effect to those principles, because the principles are not compatible with what the activity is.

The strongest objection to this argument is that it proves too much. If the principles in Article 5 are read as strict design requirements, then a wide range of legitimate processing activities, not just LLM training, would fail. Modern data analytics, longitudinal medical research, and certain forms of public administration involve forms of processing that sit uncomfortably with strict purpose limitation and data minimisation, and GDPR has not generally been read to render them unlawful. The objection has force, and it points to a real interpretive choice. The choice is between reading Article 5 as setting principles that must be substantively honoured but allow for practical accommodation in cases of demonstrable necessity, and reading it as setting procedural requirements that can be discharged through documentation and risk assessment. The literature has tended toward the second reading because the first appears too restrictive, but the second has the consequence that Article 5 places no real substantive constraint on activities that conduct themselves as though purpose limitation does not apply. The LLM case is where the second reading's emptiness becomes visible. If documentation suffices, then any activity can comply with Article 5 by documenting its non-compliance carefully enough, and the principles cease to do protective work. The structural-incompatibility argument is, at its sharpest, an argument that the second reading has been applied to LLMs in a way that empties Article 5 of content, and that the protective scheme cannot survive that emptying.

5.6 From compliance gaps to structural incompatibility

The chapter has worked through three sets of GDPR obligations whose application to LLM training the literature review had not developed in detail: Article 9 on special-category data, Articles 15 to 22 on data subject rights, and Article 5 on processing principles together with Article 25 on data protection by design. The argument across these sections has been cumulative rather than additive. Each section identifies a specific doctrinal mechanism that does not work as drafted in the LLM context, and the mechanisms do not fail for unrelated

reasons. They fail because GDPR's regulatory architecture presupposes a kind of processing relationship, bounded, purpose-specific, individually addressable, capable of being documented and controlled, that frontier-scale foundation-model training does not instantiate.

The cumulative pattern across the obligations the chapter has worked through supports a stronger conclusion than the literature review allowed. The compliance gaps are not isolated friction points but symptoms of a categorical mismatch on the doctrinal reading defended here. A reader who rejects any single one of the chapter's doctrinal commitments, in the ways 7.5 will set out, would reach different conclusions on the corresponding Article, but the structural-mismatch diagnosis survives even on more permissive readings of any individual one.

If GDPR compliance is structurally unavailable for large-scale LLM development, the harms the regulation was designed to prevent are not being prevented, and what protective scheme should apply is an open question. Section 2.4 began that question with the fundamental rights framework. The next chapter develops it by examining what protections, if any, can be reconstructed from outside the data protection regime. The doctrinal gap mapped here is the precondition for that examination.

Chapter 6. Fundamental Rights, Ethics, and Governance

6.1 Introduction

Chapter 5 ended on a doctrinal verdict. The GDPR's central obligations cannot be satisfied by frontier-scale LLM training, and the failure is not the kind that better engineering or sharper guidance will repair. That conclusion matters, but it is incomplete on its own. A regulatory regime can fail in ways that matter morally and constitutionally even if no enforcement action follows from the failure, and a fair assessment of the situation has to ask both what the regime was meant to protect and how the available alternatives are doing in protecting it.

The chapter develops the normative position the literature review pointed toward and tests the regulatory responses that have actually been produced against it. The values are what the governance has to deliver; the governance is what the values are tested against.

6.2 The Charter as a constitutional floor

The literature review introduced Articles 7 and 8 of the EU Charter of Fundamental Rights as the constitutional foundations the GDPR rests on, but did not work through the case law that gives them their actual force. The omission matters because, in EU constitutional practice, the Charter is not a rhetorical flourish above the GDPR but a binding constraint on it. Where the GDPR is interpreted in ways that systematically undercut Charter rights, that interpretation is constitutionally vulnerable, and the CJEU has on several occasions struck down secondary legislation or specific applications of it for failing to honour Charter standards. The relevance for the dissertation's argument is that the failure analysed in Chapter 5 is not merely a regulatory gap that Member States or the Commission can address through implementing acts. It is a gap that the Charter requires them to address.

The leading line of Charter jurisprudence on data protection begins with Digital Rights Ireland (Joined Cases C-293/12 and C-594/12, 2014), in which the CJEU invalidated the Data Retention Directive 2006/24/EC on the ground that its general and indiscriminate retention of telecommunications data was incompatible with Articles 7 and 8 of the Charter. The reasoning matters. The Court held that interferences with Charter rights must be limited to what is strictly necessary, must be proportionate to a legitimate objective, and must be subject

to procedural safeguards adequate to the seriousness of the interference. General and indiscriminate processing, in the Court's view, fails the proportionality test because it does not differentiate between data subjects on the basis of their relationship to the asserted public interest. The implications for LLM training are significant. Web-scraped corpus construction is general and indiscriminate by design. It does not differentiate between data subjects, does not target processing on the basis of any specific public interest, and does not implement the procedural safeguards Digital Rights Ireland identified as necessary. The pattern the case rules out is the pattern frontier LLM training fits exactly.

Schrems I (Case C-362/14, 2015) and Schrems II (Case C-311/18, 2020) extended the Charter's reach into transnational data flows and reinforced the Court's willingness to apply Charter standards even at significant economic and geopolitical cost. Schrems II in particular invalidated the EU-US Privacy Shield framework on the ground that US surveillance practices did not provide an adequate level of protection equivalent to that guaranteed by EU law, with the Charter treated as the operative standard. The Court's reasoning made clear that the Charter applies regardless of where the processing occurs and regardless of who is conducting it, so long as the data of EU subjects is involved. For the LLM context, Schrems II's significance is twofold. It confirms the Charter as the binding standard against which protective regimes are measured, and it establishes that the Court is willing to disrupt economically important arrangements where the standard is not met.

La Quadrature du Net (Joined Cases C-511/18, C-512/18 and C-520/18, 2020) is the most recent major contribution and the most directly relevant to the LLM analysis. The Court reaffirmed Digital Rights Ireland's prohibition on general and indiscriminate retention but introduced more granularity in identifying when targeted retention can be Charter-compatible. The opinion's significance for the present argument is that it treats general and indiscriminate processing of personal data as presumptively unlawful under the Charter, with the burden on the processor or the legislature to demonstrate that specific safeguards bring the processing within Charter-acceptable bounds. The mode of LLM training described in Chapter 4, corpus aggregation at internet scale, no individual-subject differentiation, no targeted procedural safeguards, indefinite retention in trained weights, does not fit any of the categories La Quadrature du Net identified as potentially Charter-compatible.

The case law supports a strong argument. If the Charter prohibits general and indiscriminate retention of telecommunications metadata, requires demonstrated necessity and

proportionality for any interference with Articles 7 and 8, and treats large-scale processing without individual differentiation as presumptively unlawful, then the question is not whether LLM training presents Charter problems but how those problems can possibly be reconciled with Charter compliance. The literature has not engaged this question seriously, in part because much of the policy debate has been conducted at the GDPR level rather than the Charter level. The Charter case law combined with the empirical analysis of Chapter 4 suggests that GDPR's failure is not the end of the legal story. It is the beginning of a constitutional story that has not yet been told.

There is a counter-argument worth registering. Charter jurisprudence developed in the context of state surveillance, and a reader sympathetic to LLM developers might argue that the case law does not transpose cleanly because the public-private distinction matters and because the data being processed is in some sense already public. The argument has surface force but does not survive examination. The Charter's scope is governed by Article 51, which binds Member States and Union institutions when implementing Union law; it does not bind private actors directly in the way it binds the state. The dissertation's argument does not require it to. The relevant route is indirect. The GDPR is a Union legal act adopted to give effect to Article 8, and Member States and Union institutions are bound by the Charter when interpreting and applying it. That is enough for Charter standards to constrain how the GDPR's lawful-basis, necessity, and proportionality requirements are read in the LLM context, regardless of whether the processor is the state or a private firm. The CJEU's case law on horizontal effect has accepted that fundamental rights can apply between private parties through secondary legislation that gives them effect, including in fields such as employment where Charter-based principles have been applied to private-sector relationships (Case C-414/16 Egenberger [2018]; Cases C-569/16 and C-570/16 Bauer and Willmeroth [2018]; Case C-68/17 IR v JQ [2018]; Case C-193/17 Cresco Investigation [2019]; Frantziou, 2019). None of this requires direct horizontal application of the Charter to LLM developers. It requires only that supervisory authorities and courts, when applying the GDPR to those developers, do so in a way that honours the Charter's prohibition on indiscriminate processing. The "already public" point is exactly the move Article 9(2)(e) makes and which Chapter 5 showed the CJEU has explicitly rejected. The Charter analysis the chapter develops is not a stretch of the case law but its natural application to a new context that the case law's reasoning straightforwardly covers.

6.3 Contextual integrity and the social meaning of privacy

The literature review used Nissenbaum's contextual integrity framework to describe what is wrong with web scraping, and engaged honestly with its limits as a guide to what to do about it. The chapter develops this further by treating contextual integrity not as one normative tool among others but as the most useful account of what privacy actually protects, against which any regulatory regime can be measured. That is a stronger claim than the literature review made, and it warrants careful argument.

The dominant alternative framings of privacy in the regulatory literature treat it either as a control right, the data subject's right to control information about themselves, or as a confidentiality interest, the right to keep information from being disclosed. Both framings have intellectual force and substantial doctrinal support, but both struggle in the LLM context for connected reasons. The control framing presupposes that data subjects can meaningfully exercise control, which the analysis in Chapter 4 has shown is not the case at scale. The confidentiality framing presupposes a clear distinction between disclosed and undisclosed information, which the LLM training context, where data was disclosed in one context and is being processed in another, collapses. Neither framing captures the structure of the harm being inflicted because neither was designed for processing that operates on already-disclosed information at population scale.

The contextual integrity framework, by contrast, explains exactly the kind of harm the LLM context produces. Privacy violations on this account are not failures of control or confidentiality but failures of contextual norm-fit. Information that has been disclosed in one context is being processed in another in ways the original context's norms would not permit. Take a medical disclosure in a patient-support forum. The disclosing person chose to share it, but only within a context whose norms permit certain flows and prohibit others. The flows the forum's norms permit are obvious enough: other forum members, clinicians who happen to be reading, researchers working under specific ethical protocols. The flows it prohibits are equally obvious: commercial extraction without consent, or integration into systems that will use the information to generate plausible content about the discloser to third parties. The training pipeline described in Chapter 4 systematically violates these norms, and that violation is the harm regardless of whether any GDPR article has been breached.

The framework's limitations need to be engaged honestly. Contextual integrity does not specify how to resolve conflicts between competing contextual norms, contexts are not static, and where an LLM synthesises content drawn from many contexts at once it is impossible to identify a single "violated context" against which a remedy might be sought. These are real limitations, but they bear on what contextual integrity can prescribe, not on what it can diagnose. They mean the framework cannot on its own generate a complete regulatory architecture; they do not mean its account of what privacy is and what privacy violations involve is wrong. The chapter uses it in the more limited but still substantive role of providing the standard against which regulatory architectures can be evaluated. A regime that allows population-scale norm violation to occur unchecked is not protecting privacy in the sense privacy actually matters, regardless of whether it satisfies its own internal compliance standards. This is the standard the rest of this chapter uses to evaluate the AI Act and other recent governance responses.

6.4 Collective harms and the limits of individual rights

The literature review identified a structural problem with the individual-rights architecture GDPR rests on, drawing on Mantelero (2022) and Wachter, Mittelstadt and Russell (2021) to argue that the framework cannot capture harms that are collective and diffuse rather than individually particularised. The chapter develops this argument into a positive claim about what an adequate regulatory regime would have to do, because the literature review only had room to identify the problem, not to develop its implications.

The mismatch can be stated directly. GDPR's protective scheme operates by giving each data subject rights they can individually exercise: access, rectification, erasure, portability, objection, the right against solely automated decision-making. These rights are individual in two senses: conferred on individual data subjects, and exercisable against individual instances of processing. The scheme presupposes every harm can be decomposed into individual rights violations a single data subject can detect, attribute, and remedy. Where that decomposition works, the scheme works. Where it does not, it systematically under-detects and under-remedies harm regardless of enforcement rigour.

LLM training produces harms whose decomposition into individual-rights violations is imperfect at best. Demographic biases encoded in trained models are one example: the harm to any single member of an affected group is small but the aggregate harm is substantial, and

no individual can demonstrate a violation enforceable on its own. Erosion of contextual norms across the open web is a second: the harm is to a shared social practice rather than to any particular person's data, and the individual-rights framework has no language for the violation. The long-term reshaping of expressive practices, as people learn what they post may be ingested by training pipelines, is a third: the harm is to potential disclosures rather than actual ones. None fits the individual-rights mould. All are real. The regulatory architecture has no mechanism for any of them.

Mantelero's (2022) collective-rights work is the most developed academic treatment of this gap, and the chapter follows him in arguing that an adequate regulatory regime would need to provide for group standing, collective remedies, and enforcement on behalf of populations rather than only on behalf of individuals. The objection that group rights are doctrinally untidy, which is the standard liberal-individualist response to collective-rights proposals, has force in some contexts but is not decisive here. Group-level protections already exist in EU law in areas like discrimination and consumer protection, and the doctrinal machinery for collective enforcement is not missing. It is just not deployed in the data protection context. The reason for that omission is partly historical, since GDPR's drafters were responding to a different generation of privacy threats, and partly political, since group rights raise difficult questions about who speaks for the group. It is not principled, and it can be addressed by regulatory design.

A deeper version of the objection deserves engaging directly. Methodological individualism, in its strict form, holds that groups do not have rights in a metaphysically robust sense: what looks like a group right is always reducible to individual rights held by individual members. The objection has serious philosophical support. The response here is partial but sufficient. The chapter accepts that the methodological-individualist position may be right about what rights metaphysically are, while denying that this settles the question being asked. The claim is not that the GDPR has failed to recognise metaphysically real group rights. It is that the GDPR's enforcement architecture, which presupposes every harm can be addressed through individually-exercisable claims by individually-identifiable data subjects, fails to detect or remedy harms whose distribution across populations defeats individual decomposition in practice. That claim is compatible with even a strict methodological individualism: a methodological individualist can consistently accept that some harms are best remedied through collective enforcement mechanisms while maintaining the underlying rights are individual. The argument is at the level of architecture, not ontology. Where it is more

contestable is at the point where it slides from "individual enforcement is inadequate" to "collective rights are required". The reform direction in 7.4 leans on the first claim, which is the safer one, and treats the second as a more ambitious version that would need separate philosophical defense.

Whether the AI Act provides the collective architecture this implies is the question the remaining sections take up.

6.5 The asymmetry problem

The ethical argument the literature review sketched through Bender et al. (2021) and Zuboff (2019) is worth restating here not as ethical critique but as a structural observation about the conditions any successor regulatory regime would have to address. Ethical critiques can be deflected; the response "but our developers care deeply about ethics" is always available. Structural observations about asymmetric power require structural responses.

The asymmetry runs along three lines that are worth keeping distinct because the regulatory responses to each are different. There is what developers know and what data subjects do not. Developers know which corpora they pulled, what filtering they ran, what the model has learned to do well. Data subjects, in most cases, do not know their data was processed at all, and the technical means to find out are not available to them even if they suspected. There is also the question of who captures the value. Training produces commercial value, that value flows to the developer and through deployment to paying users of the developer’s products, and the people whose contributions made training possible receive nothing. There is, finally, the institutional question. Developers retain legal teams, employ lobbyists, and participate in the regulatory consultations that shape the environment they then operate in. Data subjects are atomised, unrepresented as a class, and unable to coordinate at the scale the harm operates on. Some of this is a product of the technology, but the institutional layer is the part that regulation can actually adjust.

Any regulatory response that does not address the asymmetries directly will be absorbed by the activity it is meant to constrain. This is Mittelstadt’s ethics-washing argument applied to formal instruments rather than to voluntary commitments. Documentation requirements that the data subject never sees do not close the informational gap. Fines paid to the regulator do not redistribute the economic surplus the activity produces. Standing for individual data subjects does not give a population the institutional capacity to confront the firms processing

their data on aggregate. Each of these gaps is in principle correctable, but only by regulatory design that takes asymmetry as a starting point rather than treating it as a background condition the regulation operates against.

The empirical record supports the structural framing. Hong et al. (2025), already cited in the literature review, documented that web-scraped corpus pipelines routinely disregard restriction signals like robots.txt exclusion protocols. The pattern of voluntary commitments not surviving contact with commercial incentives is now well established across the field. The lesson is not that developers are uniquely bad actors but that the institutional environment selects for non-compliance with non-binding constraints, which is what asymmetric power predicts and what any successor regime must directly counter.

6.6 The EDPB framework and its visible limits

The European Data Protection Board has been the most active institutional contributor to the LLM regulatory question over the period the dissertation covers. The ChatGPT Taskforce report of May 2024 (EDPB, 2024a) and Opinion 28/2024 of December 2024 (EDPB, 2024b) together represent the most authoritative European statement on data protection in AI model development, and Section 2.6 acknowledged the genuine work these documents represent. The chapter takes for granted that the EDPB has done the legal analysis carefully. The harder question is what its framework actually does in practice.

The Taskforce report presents itself as preliminary, observing that investigations are ongoing and refraining from definitive findings on legal-basis questions. Opinion 28/2024 is more substantive but is structured around case-by-case assessment, with the controller required to demonstrate, model by model, that extraction risks are insignificant. Both documents acknowledge the practical difficulties of compliance and insist on full application of the relevant obligations, without resolving the gap between those two positions. Section 2.6 noted that this pattern, identifying a compliance gap and providing no pathway through it, generates legal uncertainty rather than compliance. The pattern has not changed in the eighteen months since the Taskforce report appeared.

Measured against the standard the chapter has been building toward, the EDPB framework has visible weaknesses across the dimensions that matter (EDPB, 2024a, 2024b). The Charter is the largest gap: Opinion 28/2024 treats GDPR application to LLMs as the only legal question on the table, and the constitutional question that Digital Rights Ireland and La

Quadrature du Net put at the centre of any analysis of indiscriminate processing is not engaged. Contextual norms appear only in a soft register, with reasonable expectations acknowledged but no specification of how they weigh against developer interests or what a population-scale violation would have to look like before consequences attached. The protective mechanism is conventionally individualist throughout, with no mechanism for population-level redress. And on asymmetry the framework loads documentation onto developers without addressing the informational, economic, or institutional gaps Section 6.5 identified.

This is not a criticism of the EDPB's good faith or its legal craft. It is an observation that the framework operates within the existing regulatory architecture rather than questioning that architecture, and that the deficiencies the dissertation has identified are properties of the architecture itself. The EDPB cannot, by issuing opinions and guidance, repair structural problems that originate in the GDPR's text and the Charter's underapplication. Reform of the kind the chapter eventually points toward would have to come from elsewhere.

6.7 The Italian Garante: case study of an enforcement model

The Italian data protection authority has been the most visible enforcer in this space, and its trajectory from 2023 through early 2026 is a useful test case for what the existing regulatory architecture can and cannot do.

In March 2023, the Garante imposed an immediate temporary block on ChatGPT in Italy, citing failures to identify a legal basis for training data processing, transparency violations, and the absence of age verification mechanisms. The block was lifted within a month after OpenAI implemented corrective measures including a privacy notice, an opt-out mechanism for training data, and age verification (Garante, 2023). The 2023 intervention is widely cited, including in the literature review, as evidence that European data protection regulators can produce concrete operational changes in major AI deployments through regulatory pressure rather than formal sanctions.

In November 2024, after a wide-ranging investigation, the Garante issued a formal decision imposing a €15 million fine on OpenAI for processing personal data without an adequate legal basis, failing to comply with transparency obligations, failing to report the March 2023 data breach, and lacking adequate age verification (Garante, 2024). The decision exercised powers under Article 166(7) of the Italian Data Protection Code to order a six-month public

information campaign across Italian media, going beyond financial penalty to require a substantive remedial measure addressing the informational asymmetry the dissertation has identified. At the time of the literature review, this decision was the most prominent enforcement action in the field.

In March 2025, the Court of Rome temporarily suspended the fine pending appeal. On 18 March 2026, the Court of Rome annulled the decision in its entirety (Court of Rome, 2026). At the time of writing, the full reasoning has not been published, and a Garante appeal remains procedurally available. The annulment may have addressed the proportionality of the fine, the substantive findings of GDPR violation, or both. What it has produced, regardless of its specific legal reasoning, is the disappearance of the only completed GDPR enforcement decision against an LLM developer in Europe for the launch period of generative AI to the public.

The implication for the present analysis is significant. Section 2.6 used the Garante's 2023 intervention to argue that regulatory engagement could produce operational change even where formal enforcement was difficult. That argument survives. The 2024 fine and its 2026 annulment illustrate something different. Formal enforcement, where it has been attempted, has not survived judicial review. Whatever the Court of Rome's specific reasoning, the practical consequence is that a developer challenging a GDPR enforcement decision in this area can reasonably hope to have the decision overturned, and that the costs of compliance with the framework as drafted may be lower in expectation than the costs of contesting it through litigation. This is not a structural feature of strong enforcement regimes.

The Garante case also illustrates the limits of the current architecture in a different way. After OpenAI established its European headquarters in Ireland in February 2024, the Irish Data Protection Commission became the lead supervisory authority under the GDPR's one-stop-shop mechanism, and pending investigations in other Member States have been transferred to or absorbed into the Irish process (EDPB, 2024a). The DPC has historically been slower and more procedurally cautious than the Garante, and the relocation of regulatory authority to a jurisdiction with weaker enforcement traditions is an example of the kind of regulatory arbitrage Section 2.6 flagged in connection with Veale and Zuiderveen Borgesius (2021). The result is that, in a structural sense, the move of OpenAI's European establishment from California to Ireland produced a quieter regulatory environment for the company, which is the opposite of what the GDPR's enforcement architecture is intended to deliver.

Tested against the standard developed earlier in this chapter, the Garante case scores poorly even on the institution's own criteria. The 2024 decision came closer than any other EU enforcement action to addressing the asymmetry problem, through the public information campaign requirement, but the decision has now been overturned. The Charter analysis was not central to the Garante's reasoning. Collective enforcement was not attempted, and the Italian courts' procedural framework does not provide for it in this context. Contextual integrity was not engaged at all. What the case demonstrates is that the most assertive single regulator in Europe, applying the existing framework with the most willingness to use formal sanctions, has not, in the period covered by the dissertation, produced a single sustained enforcement outcome against an LLM developer. The architecture is not working at the level of the activity it is meant to constrain.

6.8 The CNIL's reconstruction of legitimate interest

The French Commission Nationale de l'Informatique et des Libertés (CNIL) has taken a different approach. Rather than pursuing enforcement, the CNIL has invested heavily in producing operational guidance for AI developers, with the apparent objective of providing a practicable compliance pathway within the GDPR's existing architecture. The CNIL's recommendations on legitimate interest as a legal basis for AI training (CNIL, 2025a) and on web-scraping practices (CNIL, 2025b), both published in June 2025, are the most detailed regulatory specifications of what compliance might look like in this area. Examining them carefully shows what the existing framework can deliver when a regulator takes the trouble to make it operationally concrete, and where the limits sit.

The CNIL's legitimate-interest guidance accepts the literature review's position that Article 6(1)(f) is the most plausible legal basis for AI training in the absence of consent. The guidance specifies that the controller must define the purpose precisely, demonstrate that the processing is necessary in a strict sense, conduct and document a balancing assessment that includes both development-phase and deployment-phase risks, and implement specific safeguards including timely deletion of irrelevant data, consideration of synthetic data alternatives, and a discretionary right to object exercisable at the data subject's request (CNIL, 2025a). The web-scraping recommendations supplement this with specific technical requirements: pre-defined collection criteria, exclusion of websites that signal opposition to scraping (including through robots.txt and CAPTCHA mechanisms), exclusion of certain data

categories, restriction to freely accessible content, and publication of a list of websites scraped (CNIL, 2025b).

The guidance is, in its own terms, careful and operationally specific. It is the closest any European regulator has come to a workable compliance pathway. Tested against the standard developed earlier in this chapter, however, it shows the limits of the existing architecture even when implemented at the highest level of regulatory craft.

The Charter question goes unaddressed. The CNIL does not engage with whether the activities its guidance regulates are compatible with Articles 7 and 8 as Digital Rights Ireland and La Quadrature du Net have developed them. The mitigations the guidance specifies, exclusion of certain websites and certain data categories, are useful at the margin but do not turn web-scale corpus construction into the targeted, individually-differentiated processing the Charter case law treats as the precondition for proportionality. On contextual norms the CNIL goes further than the EDPB. The treatment of "reasonable expectations" is more substantive, and the requirement to publish a list of scraped websites is a meaningful gesture toward the disclosing context. But the operative logic is that those expectations are honoured procedurally, through opt-outs and notice, rather than substantively, through restrictions on the activity that does the violating. The collective-enforcement gap is more obvious. Opt-outs are individual, complaints are individual, the regulator's enforcement architecture is individual. The asymmetries the chapter has been tracking are partially addressed at the informational level, through documentation, and not at all at the economic or institutional levels. The CNIL's guidance is the best version of the existing framework. It falls short of the standard the chapter has been measuring against on every dimension, not because the CNIL has done anything wrong, but because the architecture it is operating within does not contain the mechanisms the standard requires.

There is a more pointed observation embedded in this. The CNIL's guidance and the EDPB's broader framework occupy the same structural position despite their differences in style: both make the existing architecture work as well as it can while leaving the architecture's limitations untouched. That work is genuinely useful for developers seeking to demonstrate good-faith compliance. It does not address whether the framework's substantive protections are being delivered.

6.9 The AI Act and the transparency template

The EU AI Act, formally Regulation 2024/1689, entered into force on 1 August 2024 with a phased application timeline that brought obligations for providers of general-purpose AI (GPAI) models into effect on 2 August 2025 (European Parliament and Council, 2024). Section 2.6 noted that the AI Act adds a parallel regulatory layer to the GDPR and identified the coordination problem this creates. The chapter takes that diagnosis as established and focuses on a specific question the literature review did not have room for: whether the AI Act's most directly relevant provision, Article 53(1)(d) on training data summaries, contributes meaningfully to the protective scheme the dissertation has been measuring against the standard developed earlier in this chapter.

Article 53(1)(d) requires providers of GPAI models to draw up and make publicly available a "sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office". The European Commission published the mandatory template in July 2025, with the obligation operative from August 2025 for new models and August 2027 for models already on the market, and AI Office enforcement, carrying fines of up to €15 million or 3% of global annual revenue, beginning in August 2026 (European Commission, 2025).

The template is structured around three information blocks: provider metadata, a listing of main data source categories, and processing and governance information on copyright, illegal content, and data protection. Providers must distinguish public, licensed, scraped, user, synthetic, and other sources, with the level of detail varying by type: public sources require names and links, private sources can be described in general terms when commercial sensitivity is invoked, and web scraping requires only a summary list of domains rather than URL-level disclosure (European Commission, 2025).

The provision is, on its own terms, a substantive transparency requirement. Measured against the standard developed earlier in this chapter, it is also conspicuously narrow. The transparency it provides is at the level of categories and provider self-reporting, not at the level of individual data subjects whose data has been included. A person whose blog post is in the training corpus learns nothing from an Article 53(1)(d) summary that they did not already know in general terms, and a medical disclosure on a patient-support forum is not addressed by a summary that says the model was trained on "scraped online content from health-related domains". The asymmetry Section 2.5 identified is largely untouched by an instrument that operates at the corpus-category level.

The provision also exempts substantive content from disclosure where commercial sensitivity is invoked. Trade secret allowances are standard regulatory practice and are not in themselves objectionable. In the LLM context, however, they have a specific effect: the data sources most likely to be commercially sensitive are also the data sources most likely to raise the substantive privacy concerns the dissertation has identified. A licensed dataset purchased from a data broker is more likely to contain individually identifiable personal data than a Common Crawl extraction, and is also more likely to be subject to confidentiality protections under the licensing agreement. The template's structure may, paradoxically, deliver less transparency where the privacy stakes are highest.

Article 53(1)(d) does not engage the Charter at all. It was drafted as a copyright transparency instrument as much as a data protection one. As a contextual norm-fit instrument it is purely descriptive: it tells the public what categories of data were used without engaging whether using them was appropriate to the contexts in which the data was originally produced. The collective dimension fares better. Enforcement runs through the AI Office at EU level rather than depending on individual data subject action, a step toward the collective architecture earlier sections identified as necessary, though the Office's enforcement record will not become testable until August 2026. On the asymmetries the provision delivers a partial gain: public disclosure modestly addresses the informational gap, while the economic and institutional gaps are untouched.

The broader concern is the one Veale and Zuiderveen Borgesius (2021) identified and Section 2.6 elaborated. The AI Act and the GDPR operate as parallel rather than integrated regimes, and the coordination mechanisms between them remain underdeveloped. A developer faces transparency obligations under the AI Act, separate transparency obligations under the GDPR's Articles 13 and 14, and CNIL-style guidance overlaid on both, without any of the three delivering the substantive protection the standard requires. Minssen, Vayena and Cohen's (2023) warning about dense regulatory coverage masking substantive gaps applies with particular force here. The picture is further complicated by the Data Act (European Parliament and Council, 2023), which adds another layer of obligations on data access, sharing, and switching that overlaps with both regimes without resolving the coordination question the AI Act and GDPR already produce.

6.10 The UK after divergence

A note on the UK is appropriate given the dissertation's institutional context. The Information Commissioner's Office has produced substantial guidance on AI and data protection (ICO, 2024), broadly aligned with the EDPB on the technical questions including memorisation and the deployed-model question, but it has not at the time of writing brought formal enforcement against a major LLM developer. The UK government's pro-innovation strategy has emphasised flexibility and sectoral application over consolidated architecture, recent UK data protection reform has not specifically addressed LLM training, and the proposed UK AI Bill announced in 2025 has not produced a binding instrument. The result is a regulatory environment formally similar to the EU's but with the enforcement layer thinned out. Measured against the standard the chapter has been using, this is worse rather than better: the standard's requirements are honoured less, because the EU architecture at least produces guidance and enforcement attempts even where those fall short. The dissertation's argument is not that the EU framework is uniquely defective and other jurisdictions do better. It is that the EU framework is the most developed example of a regulatory approach that does not fit the activity it is asked to govern, and that other jurisdictions face the same structural problem with fewer of the partial protective mechanisms the EU has at least attempted.

6.11 Conclusion

None of the institutions surveyed has done badly within the constraints they face. The shortfall is not a failure of regulatory craft. It is a failure of the architecture to be the kind of architecture that could deliver the protection the activity calls for. The constructive question Chapter 7 takes up is what kind of architecture could.

Chapter 7. Conclusions and Recommendations

7.1 Introduction

The dissertation's central thesis is that GDPR compliance is structurally rather than contingently unavailable to large-scale LLM development on web-scraped personal data, and that the failure is visible at the level of doctrine, fundamental rights, and governance practice. Better engineering, more documentation, or sharper regulatory guidance does not address the mismatch between the regulatory architecture and the activity it is being asked to regulate.

This chapter does not propose a fully worked legislative reform. The dissertation has been a critique of existing arrangements with a constructive direction, not a complete alternative framework. What the chapter does is set out the design requirements that follow from the analysis, on the basis that those requirements provide a measuring stick for any future reform proposal rather than a blueprint for one. The constructive work has to be appropriately humble, because the further work the dissertation points toward exceeds what a single-student undergraduate project can deliver.

7.2 The dissertation's argument in summary

The argument the dissertation makes, drawing on Chapters 4 to 6, can be summarised at four connected levels. Empirically, web-scraped corpora at frontier scale contain personal data and special-category data as a structural feature of the open web, and the training process does not break the link between input and model; memorisation is well-documented and exploitable, and the empirical record favours the EDPB's functional reading on which deployed models remain in scope.

Doctrinally, GDPR's central obligations cannot be satisfied in those conditions. The Article 6(1)(f) balancing test cannot be conducted with substantive content against unknown data subjects, the transparency obligations of Articles 13 and 14 cannot be met at scale, Article 9 cannot be cleared by the "manifestly made public" exemption on the strict CJEU reading, Articles 15 to 22 cannot be operationalised against a model that does not hold individual records, and Article 5 read with Article 25 cannot be honoured at the design level by an activity whose commercial logic is purpose-flexibility and data-maximisation.

Constitutionally and ethically, the protective values un-honoured when GDPR fails extend beyond GDPR's own scheme. Charter Articles 7 and 8 prohibit general and indiscriminate processing in ways no current regulatory regime engages, the harms produced are collective and diffuse in a shape the individual-rights architecture cannot reach, and the asymmetries between developers and data subjects are constitutive features of the activity rather than background conditions. The failure of governance in this space is not a failure of regulators applying the existing rules but a failure of the rules to be the kind of rules that could deliver the protection the activity requires. The constructive question is what kind of rules could.

7.3 Answering the research questions

The project plan posed three research questions. The chapter takes them in turn.

RQ1: Can GDPR's central obligations be satisfied by frontier-scale foundation-model training on web-scraped personal data?

The doctrinal analysis in Chapter 5 supports a negative answer. None of the central GDPR obligations can be satisfied at frontier scale on the doctrinal readings the dissertation defends. The answer assumes the EDPB's functional-definition reading of "personal data" prevails over the Hamburg DPA's narrow technical reading; if the Hamburg view were to prevail, deployed models would fall outside GDPR scope and a substantial portion of the analysis would not bite at the deployment stage. The dissertation has argued the EDPB's reading is doctrinally correct and consistent with CJEU jurisprudence on the functional character of GDPR concepts, but the analysis would change in important ways if the legal system settled the other way.

The answer to RQ1 is therefore no, not as drafted, on the doctrinal reading the dissertation defends. The "not as drafted" qualification matters. The obligations could in principle be satisfied by training regimes substantially different from current frontier-scale foundation-model development, including regimes that operate on smaller, consent-based, or licensed corpora. The argument is not that AI training is impossible to do compliantly, but that the specific activity of web-scale foundation-model training cannot satisfy GDPR's existing obligations.

RQ2: What protective values are being un-honoured when GDPR fails, and what is their constitutional and ethical weight?

Chapter 6's analysis supports a substantive answer. The values being un-honoured are not exhausted by the GDPR's own protective scheme. The Charter establishes Articles 7 and 8 as binding constraints whose application to LLM training has not been seriously engaged in the regulatory or academic literature, and the Charter's case law treats general and indiscriminate processing as presumptively unlawful in ways that go beyond what GDPR doctrine alone establishes. Contextual integrity supplies a positive account of what privacy protects, against which regulatory regimes can be measured, and on that account population-scale norm violation is the central harm regardless of GDPR compliance status. The collective and diffuse character of LLM-related harms places them in a category that individual-rights architecture cannot detect or remedy, even where the architecture is in principle applicable. The asymmetries between developers and data subjects are not contingent features of the present moment that better practice will repair but constitutive features of the activity that any regulatory response would have to address structurally.

The constitutional weight of these values is significant. Charter rights are not policy considerations that legislatures can balance away; they are binding standards secondary legislation must honour, and the CJEU has invalidated instruments that fail to honour them.

The answer to RQ2 is that the protective values un-honoured when GDPR fails extend significantly beyond the GDPR's own scheme, are constitutionally anchored in the Charter, and have ethical weight that does not depend on regulatory recognition. The failure of GDPR is not the failure of the entire protective scheme; it is the failure of the most prominent operative mechanism within a broader scheme that includes constitutional and ethical dimensions the regulatory debate has tended to sideline.

RQ3: What would adequate governance of LLM training on web-scraped personal data have to look like?

Chapter 6's analysis, combined with the standard the chapter develops, supports a structured answer. Adequate governance would have to honour the Charter's prohibition on general and indiscriminate processing, treat contextual norms as a substantive standard rather than as soft preferences, provide collective enforcement mechanisms appropriate to harms whose individual decomposition is imperfect, and structurally address the informational, economic, and institutional asymmetries the activity rests on. None of the current regulatory responses

examined there meets this standard, and the reasons for the shortfall are structural rather than reflective of failures of regulatory craft.

The answer to RQ3 is the substance of Section 7.4 below, which sets out the design requirements that follow from this analysis.

7.4 What reform would have to do

The literature review's closing section sketched a four-part reform direction: mandatory DPIAs with third-party audits, representative complaints mechanisms, collective redress schemes, and GDPR/AI Act coordination. The analytical chapters support that direction and develop it further. What follows is not a numbered list of requirements but the connected substance of what reform would have to engage with for any of those individual moves to do real protective work.

The Charter is the obvious starting point because it sets a standard that no current regulatory response is even attempting to meet. Articles 7 and 8 as developed in Digital Rights Ireland and La Quadrature du Net prohibit general and indiscriminate processing of personal data, and frontier-scale web scraping is general and indiscriminate by design. The reform options are narrower than they look. Either the activity has to be reconciled with the Charter through some demonstrable showing of necessity and proportionality, or the activity has to change so that it falls within the targeted, individually-differentiated processing the case law treats as permissible. I do not take a view on which of these is achievable. I do take a view on the constitutional question being one that has to be confronted rather than left aside, and the EDPB framework as currently constituted does not confront it. Connected to this is what an adequate regime would have to do with contextual norms. The framework Chapter 6 developed treats privacy violations as failures of contextual norm-fit, and a successor regime would treat such failures as grounds for substantive restriction rather than as procedural problems to be managed through opt-outs and notice. The CNIL's June 2025 guidance is the closest a European regulator has come to taking contextual norms seriously, and it does not go this far. Going further would mean restricting the use of categories of data drawn from contexts whose norms the activity violates, regardless of whether the data happens to be technically public and regardless of whether individual data subjects have exercised opt-outs. This is the move the existing framework has consistently declined to make, and it is the substantive expression of what the Charter analysis already demands.

The enforcement architecture is where the GDPR's individual-rights model cannot do the work alone. What is needed is enforcement at the level of populations: standing for civil society organisations and designated representative bodies, class-wide remedies, and regulatory action that does not depend on identifying a complainant whose specific rights were violated to a documentable degree. The doctrinal machinery already exists in EU consumer protection and discrimination law. It has not been transposed to data protection, partly for historical reasons and partly because group rights remain politically uncomfortable, but neither reason is principled and the mechanisms can be built. Enforcement reform on its own does not address the deeper problem the analysis identified, which is that asymmetry is constitutive of the activity rather than peripheral to it. Documentation requirements that the data subject never sees, fines paid to the regulator, and procedural opt-outs do not redistribute anything; they manage the activity without changing the institutional environment in which it operates. What would change that environment is harder. Mandatory revenue-sharing with affected populations is one option. Data trusts or fiduciary structures that vest representation in independent bodies acting for data subjects is another. Compute-access provisions that distribute the benefits of training beyond the developer firms are a third. Each is a substantial regulatory move with implications the dissertation has not worked through. The structural point is that governance which leaves the institutional environment untouched will be absorbed by the activity it is supposed to constrain.

Of the three options the previous paragraph names, data trusts are the one most worth developing. Two reasons. The existing UK and European policy work on them is more advanced than for the others, and the institutional architecture they propose maps closely onto the asymmetry problem the analysis has identified. The basic idea, developed across UK and European policy work on data intermediaries over recent years, is that a data trust is a legal structure in which independent trustees hold representation rights for a population of data subjects, with a fiduciary duty to act in that population's interest. The trustees can negotiate with developers on behalf of the population, set conditions for data use, and enforce those conditions through mechanisms the population could not deploy individually. The structure does not require methodologically robust group rights, which is why it survives the objection Section 6.4 engaged: the fiduciary duty runs to individuals collectively rather than to the group as a metaphysical entity, and the trust is a coordination mechanism rather than a rights-bearer in its own right.

The proposal has more legal grounding than is sometimes recognised. Article 80 GDPR already permits not-for-profit bodies to represent data subjects in lodging complaints and exercising rights, and Article 80(2) lets Member States extend this to enable representative actions independent of any individual mandate. The Article 80 architecture is currently underused, but EU collective redress mechanisms have been extended across consumer law in recent years through the Representative Actions Directive (European Parliament and Council, 2020), and there is no principled reason data protection should remain outside that direction of travel. A data trusts regime built on Article 80(2) would not require new primary legislation; it would require Member State implementation that designates appropriate trustee bodies and confers on them the standing the article already contemplates. The AI Act's notification mechanisms could provide a parallel route, particularly through the AI Office's power to designate civil society representatives in its consultation processes, which could be developed into more substantive enforcement standing over time.

For the proposal to bite at the level of the activity, the trustees need technical mechanisms to verify what they are negotiating about. The institutional design above does not by itself give a trustee any way to check whether a developer's claimed dataset composition matches what was actually used, or whether a model has memorised the kinds of content the trust is meant to protect against. Three technical primitives would have to be in place. The first is dataset provenance tracking: cryptographic commitments published by developers at training time that bind a model to the corpus it was trained on, so that subsequent claims about composition are auditable rather than self-reported. Common Crawl already publishes manifests of its snapshots, and provenance schemes like C2PA for media are being extended to text corpora; the technical building blocks exist. The second is a training data registry, structured analogously to the AI Act's Article 53(1)(d) summary template but with sufficient granularity (domain-level rather than category-level) for trustees to query. The third is a third-party audit interface that lets trustees probe deployed models for memorisation of content drawn from the represented population, without requiring full white-box access to the weights. The technical literature on differential-privacy auditing and membership inference attacks (Carlini et al., 2023) provides the methods; what is missing is the institutional requirement that developers expose the inference endpoints in a form trustees can use. None of these primitives is technically novel. What is novel is requiring them as conditions of operating in the European market.

The proposal should also be tested against its plausible failure modes rather than presented as if it would obviously succeed. The most direct trade-off is that the audit and provenance requirements impose real engineering costs, and the firms able to bear those costs are the same large incumbents the proposal is meant to constrain. Smaller European AI firms could be disproportionately disadvantaged, and the regime could entrench rather than disrupt the existing competitive structure. The provenance commitments themselves create a secondary privacy problem, because granular dataset manifests can leak information about which individuals contributed which content, particularly where the contributing population is small. Enforceability at scale is another constraint. Trustees with standing and resources comparable to a single major regulator could not audit every model deployed in the European market, and the regime would in practice operate by sampling and high-profile interventions rather than by comprehensive coverage. Each of these is a real concern. None is a fatal objection. The relevant comparison is not between the proposal and a regime that does not face these problems, but between the proposal and the existing framework that does not address the structural asymmetries at all. Whether the trade-off is worth making is a political question the dissertation cannot settle, but it is a more honest political question than the one the existing framework currently asks.

The hardest question the proposal faces is who funds the trustee bodies and how their independence is preserved. State funding through national data protection authorities raises capture risks at the regulator level. Industry funding raises capture risks at the trustee level. The most plausible answer is a hybrid: a small statutory levy on the largest data-processing firms, modelled loosely on the funding mechanisms used for ombudsman services in financial services and telecoms, paid into an independent fund with allocation overseen by a body the regulators do not directly control. This is not a complete answer; it leaves real questions about who appoints the appointers, how the trustees are accountable, and what happens when the trustees and the represented population disagree. But it is concrete enough to be discussed in the policy literature, which is what a developed proposal needs to be. What it does that none of the existing regulatory responses does is address the institutional asymmetry directly, by giving the represented population a coordinating mechanism with standing, capacity, and resources commensurate with what developers already have. The structural argument the chapter has been making is that governance which leaves the institutional environment untouched will be absorbed by the activity it constrains. A data trusts regime is the version of that proposition that has the clearest institutional architecture

and the most developed policy track record. It is not the only worked proposal that could meet the standard, and the dissertation does not claim that it would by itself solve the problem. What it claims is that this is what a serious worked proposal looks like, and that any successor to the existing framework would need to engage with proposals at this level of institutional specificity rather than continuing to issue guidance and documentation requirements that the activity has demonstrated it can absorb without changing.

Sitting underneath all of this is the coordination problem between the GDPR and the AI Act. The two regimes run in parallel with thin coordination, and the result Veale and Zuiderveen Borgesius (2021) predicted has materialised: dense regulatory coverage with substantive gaps, and developers able to choose which instrument to demonstrate compliance against for any given purpose. Closing this would require either a single integrated framework or a substantially deepened coordination architecture, with clear allocation of competences between data protection authorities, the AI Office, and any sectoral bodies that emerge, and with mechanisms that ensure compliance with one regime cannot discharge obligations under another. The diagnosis has been available for years. The legislative work to act on it has not begun. Taken together, what the analysis points to is demanding and not exhaustive. Meeting these requirements would not by itself solve the problem. But not meeting them certainly would not either.

7.5 Limitations of the analysis

The dissertation's argument is open to several lines of criticism that are worth registering honestly.

The doctrinal reading defended in Chapter 5 makes substantial commitments that the legal system has not yet settled. The EDPB's functional-definition reading on the deployed-model question is supported by the empirical record but the Hamburg DPA's narrow technical reading has not been judicially overruled. The strict CJEU reading of Article 9(2)(e) was developed in the search engine context of GC and Others, and a court might accept that the LLM context is sufficiently different to warrant a more permissive reading despite the analysis I provided to the contrary. The Article 5 structural-incompatibility argument depends on rejecting a procedural reading of the principles that has substantial support in the existing literature, and the dissertation's response to the "proves too much" objection in Section 5.5 is not the only available one. If any of these readings fails in court, parts of the argument would

have to give way. The dissertation has tried to flag where each commitment sits and to acknowledge the alternatives, but the reader should treat the analysis as a defended position rather than as a settled reading.

The Charter analysis in Chapter 6 transposes case law from the state-surveillance context to the commercial AI context, and the chapter argued the transposition is principled and the case law's reasoning supports it. A reader sympathetic to LLM developers might argue that Charter doctrine in the LLM context will develop in ways the analysis has not anticipated. Section 6.2 engaged the public-private distinction by routing the Charter argument through indirect horizontal effect, with the Charter binding Member States and Union institutions when applying the GDPR to private actors rather than binding LLM developers directly, and the response avoids the strongest version of the public-private objection. What it does not do is engage the academic literature on Charter horizontal effect (Frantziou, 2019) at the depth a constitutional law specialist would. A more developed treatment would have to address competing readings of the relevant CJEU jurisprudence and the wider scholarly debate on how Charter-based principles apply between private parties. The argument is defensible at the level the dissertation makes it, but a fuller doctrinal defense would need work this dissertation does not have room for.

The collective-harms analysis in Chapter 6 follows Mantelero (2022) and Wachter, Mittelstadt and Russell (2021) in identifying a structural mismatch between individual-rights architecture and the actual shape of LLM-related harms. Section 6.4 engages the strict methodological-individualist objection partially, by arguing that the dissertation's claim is about enforcement architecture rather than the metaphysics of rights, and is therefore compatible with even a strict methodological-individualist position. The response is partial because where the chapter shifts from "individual enforcement is inadequate" to "collective rights are required", the methodological-individualist objection still has force, and a fuller treatment would need to engage the philosophical literature on group rights at greater length than an undergraduate dissertation can sustain. A reader committed to methodological individualism in rights theory would still find the chapter's treatment incomplete on the more ambitious version of the claim, even if persuaded on the safer architectural version.

The governance assessment in Chapter 6 draws on regulatory developments through early 2026, with the Court of Rome's annulment of the Garante's fine occurring six weeks before the dissertation's submission. The Court's full reasoning has not been published at the time of

writing. If the Court's reasoning, when published, addresses only the proportionality of the fine rather than the substantive findings of GDPR violation, the chapter's reading of the annulment as undermining the enforcement architecture would need softening. The dissertation has tried to indicate this uncertainty in Section 6.7, but a reader should be aware that the section's conclusions are vulnerable to subsequent regulatory developments.

Methodological limitations were acknowledged in the project plan and remain operative. The dissertation has not conducted original empirical research on training corpora or memorisation, relying instead on the existing empirical literature. It has not engaged regulators, developers, or affected communities directly. It has not modelled the economic implications of the reform direction it points toward. The data trusts proposal developed in 7.4 is concrete enough to be discussed in the policy literature, but it is not a worked policy proposal in the sense a government department would produce; the funding architecture, governance structure, and accountability mechanisms would all need substantial further development. The other reform options the chapter names, mandatory revenue-sharing and compute-access provisions, are sketched rather than developed, and a fuller treatment of the question would require all of these and more.

These limitations are real but they do not undermine the central thesis. They constrain the strength of the constructive claims the dissertation makes and identify directions for further work.

7.6 Contribution and further work

The dissertation's contribution is integrative rather than empirical or doctrinal in isolation. The empirical work on LLM training, the doctrinal work on GDPR obligations, and the normative work on fundamental rights and ethics have been conducted in adjacent parts of the academic literature, but the connections between them have not been systematically developed. The dissertation has tried to treat these three domains as a single analysis, with the empirical conditions establishing what compliance actually requires, the doctrinal analysis establishing what the regulatory framework demands, and the normative analysis establishing whether either, even if achieved, would deliver the protections at stake.

The further work the analysis points toward exceeds what an undergraduate dissertation can deliver. The constitutional question Chapter 6 raised needs more work than I could give it here, ideally a treatment that engages the EU constitutional law literature in depth. The

collective-rights analysis the same chapter sketched needs sustained engagement with the philosophical literature on group rights and comparative analysis with adjacent EU law areas where group-level protections operate. The reform direction Section 7.4 set out would require legal-economic analysis of the asymmetry-intervention proposals, which is a substantial research programme rather than a chapter. The empirical question of whether the Hamburg DPA's reading or the EDPB's reading prevails over time will be answered, if at all, by litigation rather than scholarship, and the dissertation can only register the question rather than settle it.

These are not gaps the dissertation has failed to fill but directions for further work that the dissertation's argument identifies as worth pursuing. Locating them clearly is itself a contribution: it gives subsequent work a starting point rather than a blank page.

7.7 Closing

The GDPR is not the right tool for governing frontier-scale foundation-model training on web-scraped personal data. That is not a claim that the GDPR is a bad regulation, or that data protection is the wrong frame. The GDPR is a careful instrument designed for a world of bounded, purpose-specific data processing relationships, and within that world it does real protective work. The activity I have examined is not in that world. Either the architecture changes to fit the activity, or the activity changes to fit the architecture, or some combination of the two. What cannot continue is the present arrangement, where a regulatory framework designed for one form of processing is asked to govern another and produces dense documentation and modest disclosure without delivering the protection the framework was meant to deliver.

The further claim is more modest. The materials to do better are already on the table. The Charter case law is there, the contextual integrity literature is there, collective-rights doctrine in adjacent areas of EU law is there, and the careful technical work the EDPB and the CNIL have already produced is there. The pieces of an adequate response exist. Assembling them into a coherent regime is work that exceeds what an undergraduate dissertation can do. Pointing toward the work, and saying clearly why it is needed, is what this one has tried to do.

Consolidated References

The reference list below brings together every source cited across the literature review, methodology, and analytical chapters of this dissertation. Entries are presented in alphabetical order following Harvard referencing conventions. Where a regulatory document or court case is cited, the citation follows the conventional legal format alongside the Harvard apparatus.

Bender, E., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). New York: ACM, pp. 610-623. Available at: https://doi.org/10.1145/3442188.3445922 (Accessed: 10 March 2026).

Birhane, A. and Prabhu, V.U. (2021) 'Multimodal datasets: misogyny, pornography, and malignant stereotypes', arXiv preprint arXiv:2110.01963.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D. and Papernot, N. (2021) 'Machine unlearning', in 2021 IEEE Symposium on Security and Privacy (SP). San Francisco, CA: IEEE, pp. 141-159.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A. and Raffel, C. (2021) 'Extracting Training Data from Large Language Models', in 30th USENIX Security Symposium (USENIX Security '21), pp. 2633-2650.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F. and Zhang, C. (2023) 'Quantifying Memorization Across Neural Language Models', in Proceedings of the 11th International Conference on Learning Representations (ICLR 2023).

Case C-68/17 IR v JQ [2018] ECLI:EU:C:2018:696.

Case C-136/17 GC and Others v CNIL [2019] ECLI:EU:C:2019:773.

Case C-193/17 Cresco Investigation GmbH v Markus Achatzi [2019] ECLI:EU:C:2019:43.

Case C-311/18 Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems (Schrems II) [2020] ECLI:EU:C:2020:559.

Case C-362/14 Maximillian Schrems v Data Protection Commissioner (Schrems I) [2015] ECLI:EU:C:2015:650.

Case C-414/16 Vera Egenberger v Evangelisches Werk für Diakonie und Entwicklung eV [2018] ECLI:EU:C:2018:257.

Cases C-569/16 and C-570/16 Stadt Wuppertal v Maria Elisabeth Bauer and Volker Willmeroth v Martina Broßonn [2018] ECLI:EU:C:2018:871.

CNIL (Commission Nationale de l'Informatique et des Libertés) (2025a) Recommendation: Relying on the legal basis of legitimate interests to develop an AI system. 19 June 2025. Paris: CNIL.

CNIL (Commission Nationale de l'Informatique et des Libertés) (2025b) Recommendation: The legal basis of legitimate interest, focus sheet on measures to be taken in the event of data collection through web scraping. 19 June 2025. Paris: CNIL.

Common Crawl Foundation (2024) About Common Crawl. Available at: https://commoncrawl.org/ (Accessed: April 2026).

Court of Rome (2026) Judgment no. 4153/2026 (R.G. 2025) annulling Garante per la Protezione dei Dati Personali Decision No. 755 of 2 November 2024. 18 March 2026.

Datatilsynet (2023) Generative kunstig intelligens (GenAI) [Generative Artificial Intelligence]. Copenhagen: Danish Data Protection Authority.

Dodge, J., Sap, M., Marasoćvic, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M. and Gardner, M. (2021) 'Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus', in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, pp. 1286-1305.

European Commission (2025) Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. 24 July 2025. Brussels: European AI Office.

European Data Protection Board (2024a) Report of the work undertaken by the ChatGPT Taskforce. 23 May 2024. Brussels: EDPB.

European Data Protection Board (2024b) Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. Adopted 17 December 2024. Brussels: EDPB.

European Data Protection Supervisor (2023) Large Language Models (LLM): TechSonar briefing. Brussels: EDPS.

European Parliament and Council (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, pp. 1–88.

European Parliament and Council (2020) Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of consumers and repealing Directive 2009/22/EC. OJ L 409, 4.12.2020, pp. 1–27.

European Parliament and Council (2023) Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). OJ L, 22.12.2023.

European Parliament and Council (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). OJ L, 12 July 2024.

European Union (2012) Charter of Fundamental Rights of the European Union. OJ C 326, 26.10.2012, pp. 391–407.

Frantziou, E. (2019) The Horizontal Effect of Fundamental Rights in the European Union: A Constitutional Analysis. Oxford: Oxford University Press.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S. and Leahy, C. (2021) ‘The Pile: An 800GB dataset of diverse text for language modeling’, arXiv preprint arXiv:2101.00027.

Garante per la Protezione dei Dati Personali (2023) Provvedimento del 30 marzo 2023, Limitazione provvisoria del trattamento dei dati degli utenti italiani nei confronti di OpenAI L.L.C. [Order of 30 March 2023]. Rome: Garante.

Garante per la Protezione dei Dati Personali (2024) Provvedimento del 2 novembre 2024, OpenAI L.L.C., Decision No. 755 (Doc. web n. 10085455). 2 November 2024. Rome: Garante.

Hamburg DPA (Hamburg Commissioner for Data Protection and Freedom of Information) (2024) Discussion paper: Large Language Models and Personal Data. 15 July 2024. Hamburg: HmbBfDI.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O. and Sifre, L. (2022) 'Training compute-optimal large language models', in Advances in Neural Information Processing Systems 35 (NeurIPS 2022). New Orleans, LA, pp. 30016-30030.

Hong, R., Hutson, J., Agnew, W., Huda, I., Kohno, T. and Morgenstern, J. (2025) 'A common pool of privacy problems: Legal and technical lessons from a large-scale web-scraped machine learning dataset', arXiv preprint arXiv:2506.17185. [Preprint, not peer-reviewed at time of writing.]

IAPP (International Association of Privacy Professionals) (2024) 'Do LLMs "store" personal data? This is asking the wrong question', IAPP News, 24 July.

Information Commissioner's Office (ICO) (2023) The lawful basis for web scraping to train generative AI models. London: ICO.

Information Commissioner's Office (ICO) (2024) Guidance on AI and data protection. London: ICO.

Joined Cases C-293/12 and C-594/12 Digital Rights Ireland Ltd v Minister for Communications and Kärntner Landesregierung [2014] ECLI:EU:C:2014:238.

Joined Cases C-511/18, C-512/18 and C-520/18 La Quadrature du Net and Others v Premier ministre and Others [2020] ECLI:EU:C:2020:791.

Kamarinou, D., Millard, C. and Singh, J. (2017) 'Machine Learning with Personal Data', in Leenes, R., van Brakel, R., Gutwirth, S. and de Hert, P. (eds.) Data Protection and Privacy: The Age of Intelligent Machines. Oxford: Hart Publishing, pp. 89-114.

Kuziemski, M. and Misuraca, G. (2020) 'AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings', Telecommunications Policy, 44(6), 101976.

Mantelero, A. (2022) Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI. The Hague: T.M.C. Asser Press.

Minssen, T., Vayena, E. and Cohen, I.G. (2023) 'The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models', JAMA, 330(4), pp. 315-316.

Mittelstadt, B. (2019) 'Principles alone cannot guarantee ethical AI', Nature Machine Intelligence, 1(11), pp. 501-507.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F., Ippolito, D., Choquette-Choo, C.A., Wallace, E., Tramèr, F. and Lee, K. (2023) 'Scalable extraction of training data from (production) language models', arXiv preprint arXiv:2311.17035.

Nguyen, T.T., Huynh, T.T., Nguyen, P.L., Liew, A.W.-C., Yin, H. and Nguyen, Q.V.H. (2022) 'A survey of machine unlearning', arXiv preprint arXiv:2209.02299.

Nissenbaum, H. (2019) 'Contextual Integrity Up and Down the Data Food Chain', Theoretical Inquiries in Law, 20(1), pp. 221-256.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E. and Launay, J. (2023) 'The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only', arXiv preprint arXiv:2306.01116.

Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C., von Werra, L. and Wolf, T. (2024) 'The FineWeb datasets: Decanting the web for the finest text data at scale', arXiv preprint arXiv:2406.17557.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J. (2020) 'Exploring the limits of transfer learning with a unified text-to-text transformer', Journal of Machine Learning Research, 21(140), pp. 1-67.

Soldaini, L., Kinney, R., Bhagia, A. et al. (2024) 'Dolma: An open corpus of three trillion tokens for language model pretraining research', arXiv preprint arXiv:2402.00159.

Solove, D. (2021) 'The Myth of the Privacy Paradox', George Washington Law Review, 89(1), pp. 1-51.

Thiel, D. (2023) Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford: Stanford Internet Observatory.

Veale, M. and Zuiderveen Borgesius, F. (2021) 'Demystifying the draft EU Artificial Intelligence Act', Computer Law Review International, 22(4), pp. 97-112.

Wachter, S., Mittelstadt, B. and Russell, C. (2021) 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI', Computer Law and Security Review, 41, 105567.

Zuboff, S. (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs.