PETLP: A Privacy-by-Design Pipeline for Social Media Data in AI Research

Nick Oh[1], **Giorgos D. Vrakas**[2], **Siân J.M. Brooke**[3,4], **Sasha Morinière**[5†], **Toju Duke**[6]
[1]socius labs
[2]University of Cyprus
[3]University of Amsterdam
[4]London School of Economics and Political Science
[5]Conspiracy Watch
[6]Bedrock AI
nick.sh.oh@socius.org, vrakas.giorgos-dimitris@ucy.ac.cy, s.j.m.brooke@uva.nl, toju@bedrockai.org

Abstract

Social media data presents AI researchers with overlapping obligations under the GDPR, copyright law, and platform terms – yet existing frameworks fail to integrate these regulatory domains, leaving researchers without unified guidance. We introduce **PETLP** (**P**rivacy-by-design **E**xtract, **T**ransform, **L**oad, and **P**resent), a compliance framework that embeds legal safeguards directly into extended ETL pipelines. Central to PETLP is treating Data Protection Impact Assessments as living documents that evolve from preregistration through dissemination. Through systematic Reddit analysis, we demonstrate how extraction rights fundamentally differ between qualifying research organisations (who can invoke DSM Article 3 to override platform restrictions) and commercial entities (bound by terms of service), whilst GDPR obligations apply universally. We demonstrate why true anonymisation remains unachievable for social media data and expose the legal gap between permitted dataset creation and uncertain model distribution. By structuring compliance decisions into practical workflows and simplifying institutional data management plans, PETLP enables researchers to navigate regulatory complexity with confidence, bridging the gap between legal requirements and research practice.

Decision Trees and Appendices —
https://www.arxiv.org/abs/2508.09232

1 Introduction

Social media platforms have become essential data sources for computational and social science research, enabling investigations into political movements (Solovev and Pröllochs 2022; Iqbal et al. 2022), labour inequalities (Xu 2024), discrimination (Nesterov, Hollink, and van Ossenbruggen 2024) and cultural phenomena (Nguyen et al. 2023). These platforms provide datasets of unprecedented scale and immediacy, offering insights into societal trends and human behaviour that surpass traditional methods (Bundtzen 2023). Concurrently, methodological advances – from BERT-based models (Ji et al. 2020; Ananthakrishnan et al. 2022) to Large Language Models (LLMs) (Deng et al. 2023; Alhamed, Ive, and Specia 2024; Vuruma et al. 2024) – have expanded analytical possibilities, yet with a cost of intensifying ethical and regulatory challenges.

The research landscape presents AI researchers with multiple layers of complexity. At the foundational level, the definition of social media itself continues to evolve – conventional websites now incorporate social features like user profiles, messaging, and content-sharing to enhance engagement (Garcia-Pueyo et al. 2023; He et al. 2024), blurring boundaries and creating ambiguity about applicable legal and ethical frameworks (Terzis, Veale, and Gaumann 2024). Regulatory frameworks add a second layer of uncertainty: the GDPR offers limited guidance on platform-to-researcher data sharing (European Digital Media Observatory 2022) while employing broad, often undefined concepts such as 'personal data' and 'scientific research'.

These definitional and regulatory ambiguities intersect with a third layer – platform governance tensions. Platforms increasingly invoke privacy concerns to restrict data access (Bruns 2021; Morten, Nicholas, and Viljoen 2024), despite legal precedents establishing no reasonable expectation of privacy for publicly accessible content (Rom 2010). Nevertheless, the ethical landscape presents equally significant challenges, with empirical studies demonstrating that fewer than 10% of social media research publications address ethical implications beyond securing basic Institutional Review Board (IRB) approval (Taylor and Pagliari 2018; Chiauzzi and Wicks 2019). The intersection of these three layers – evolving platform definitions, regulatory gaps, and increasingly restrictive platform policies – creates a fragmented compliance landscape where researchers face contradictory obligations that ultimately compromise both research quality and public confidence (European Digital Media Observatory 2022).

To navigate these challenges, we propose the *Privacy-by-design Extract, Transform, Load and Present (PETLP)* framework. Unlike conventional data pipelines that treat compliance as an afterthought, PETLP directly addresses each layer of complexity: it clarifies definitional ambiguities, operationalises vague regulatory concepts via living DPIAs, navigates platform restrictions through legally-grounded alternatives, and elevates ethical practice beyond checkbox compliance. The framework integrates three intersecting legal domains – data protection regulations, intellec-

[†]Work conducted during affiliation with the Open Data Institute

tual property (IP) rights, and platform contract law – transforming regulatory requirements from external constraints into structured technical workflows. PETLP addresses the fundamental question: *How can researchers responsibly access, process, and present social media data and derivative models whilst fulfilling legal obligations throughout the data pipeline?*

2 Background and Scope

Social media data presents unique challenges for AI research. Its distinctive characteristics, including mixed public-private boundaries, embedded personal information, and complex ownership structures, necessitate careful legal consideration before collection, processing, or model deployment. Understanding the legal frameworks governing such data is therefore essential for responsible AI research practice.

Our analysis examines three intersecting legal domains that collectively govern AI researchers' use of social media data: data protection, IP, and contract law. We centre this analysis on European Union law, which provides a comprehensive, principle-based governance model (Lynskey 2015; European Commission 2021). The GDPR's extraterritorial reach – applying to any processing of EU-located individuals' data regardless of researchers' location – makes compliance virtually certain for social media AI research, as global platforms inevitably include EU users' content. This broad applicability establishes EU law as the de facto baseline for international research standards.

Whilst focusing on current requirements, we acknowledge the EU AI Act (effective August 2025) introduces complementary development-phase obligations that reinforce transparency and risk management principles (see Appendix C). Researchers outside Europe should consult jurisdiction-specific guidance, such as Brown et al. (2024) for US web scraping or the AI Governance and Regulatory Archive (AGORA) (Arnold et al. 2024) for cross-jurisdictional analysis. However, GDPR's extraterritorial application means most social media AI research requires European compliance regardless of researcher location, making our framework broadly applicable.

2.1 Defining Social Media Data

Social media data refers to content generated and collected from users and their interactions within platforms, often in the context of *social relationships*[1] (Olteanu et al. 2019; Stieglitz et al. 2018; Batrinca and Treleaven 2015). This encompasses traditional social networking sites like Facebook (now Meta) and Twitter (now X), alongside content-sharing platforms (YouTube, Instagram), messaging applications with social features (WhatsApp, WeChat), and community-based platforms (Reddit, Discord) (Boyd and Ellison 2007; Obar and Wildman 2015).

This social dimension creates unique legal and ethical challenges, as individuals can be identified not merely through their own posts but through their networks of connections, interaction patterns, and community memberships – even when traditional identifiers are removed. A user's participation in specific communities, commenting patterns, or social graph can serve as digital fingerprints, making anonymisation particularly difficult and raising the stakes for privacy protection (Qian et al. 2019; Jiang et al. 2018; Tian et al. 2018).

2.2 Legal Foundations for Social Media Data Pipelines

Utilising social media data for AI training requires researchers to traverse a landscape marked by fundamental legal uncertainty (Longpre et al. 2024; Dornis and Stober 2025). Three distinct yet interconnected legal regimes govern this terrain, each imposing specific obligations that shape every stage of the research pipeline. Understanding these foundations is essential – not merely for compliance, but for designing research that can withstand scrutiny, maintain access to data sources, and preserve public trust.

The GDPR establishes comprehensive privacy principles yet provides limited operational guidance for research contexts. Key concepts remain broadly defined (European Digital Media Observatory 2022): what constitutes 'research purposes', how 'personal data' extends in social media contexts, or which activities qualify as 'scientific research'. Similarly, IP law presents ambiguity – whilst the EU's Text and Data Mining (TDM) exceptions permit researchers to extract patterns from copyright protected data, it remains unclear whether these protections extend to training AI models. Contract law adds further complexity through platform Terms of Service (ToS) that often prohibit automated access, redistribution, or AI training, creating potential conflicts with statutory research rights.

These overlapping regimes create compound risks. A single research project will typically navigate data subject rights under the GDPR, copyright protections on user-generated content, database rights held by platforms, and contractual restrictions imposed through ToS. The following sections examine each legal domain in detail, establishing the conceptual foundations necessary for understanding their technical implications and practical intersections throughout the research pipeline.

GDPR The GDPR establishes fundamental principles for data handling that directly shape how AI researchers can work with social media data. Its definition of 'processing' encompasses virtually every research activity – collecting posts, storing usernames, cleaning datasets, training models, and publishing results all constitute 'processing' under the regulation. This means GDPR compliance is not a one-time checkbox but an ongoing obligation throughout the research pipeline, from initial data collection to final model deployment.

Three overarching principles guide GDPR obligations.

[1]The European Data Protection Board (EDPB) defines social media as 'online platforms that enable the development of networks and communities of users, among which information and content is shared', where key characteristics include account creation, content sharing, and user connections (European Data Protection Supervisor 2021, p. 4).

First, *necessity and proportionality* require that data processing be both essential for research objectives and balanced against privacy impacts. Researchers cannot scrape entire subreddits *just in case* some posts prove useful – each data point must serve the stated research purpose. Second, *accountability* demands that researchers not only comply with the GDPR but demonstrate compliance through documentation and safeguards. What constitutes 'appropriate' safeguards depends on risk: analysing public product reviews requires basic security measures, while processing mental health discussions demands careful handling. Third, *rights and freedoms* of data subjects remain paramount – individuals retain rights to access, correct, or object to processing of their data, though Article 89 allows some limitations for scientific research where these rights would 'seriously impair' research objectives. Article 5 translates these into six specific principles, and Article 25 mandates that these protections be embedded into research design from the outset – *privacy by design, not by afterthought*.

The following discussion introduces four core GDPR considerations for social media research: the personal data status of social media content, controllership determination, impact assessment requirements, and legal bases for processing. Appendix A supplements this overview, providing the underlying legal rationale, regulatory interpretations, and citations that substantiate these requirements.

Social Media Data as Personal Data Social media data qualifies as personal data under the GDPR, irrespective of public accessibility. The determining factor lies not in the data's availability but in its potential to identify individuals directly or indirectly (Article 29 Data Protection Working Party 2017; Court of Justice of the European Union 2024b). Under the GDPR, data is rendered anonymous *only* when individuals cannot be identified by any means 'reasonably likely to be used' (Recital 26), accounting for future technological advances and data linkage possibilities – a standard rarely achievable with social media's rich behavioural patterns. Moreover, social media data frequently enables inference of special category data under Article 9; recent Court of Justice of the European Union (CJEU) rulings confirm that the mere possibility of inferring protected attributes activates heightened safeguards *regardless of accuracy* (Court of Justice of the European Union 2022, 2023b,a). This is particularly significant for AI research, where models may inadvertently memorise training data or develop unforeseen inference capabilities (European Data Protection Board 2024b), demanding researchers account for both intended outcomes and latent model capacities.

The Researcher as Joint Controller While research institutions formally act as controllers, controllership is functionally determined by actual influence over data processing decisions (European Data Protection Board 2020). AI researchers directly determine collection methodologies, transformation techniques, and model training approaches – decisions integral to defining the *means* of processing under Article 4(7). This paper therefore advocates viewing researchers as joint controllers with their institutions, recognising that joint controllership emerges when parties jointly determine purposes and means through complementary decisions (European Data Protection Board 2020). This perspective ensures privacy considerations permeate the entire research pipeline rather than remaining abstract institutional obligations, creating accountability where data handling decisions are actually made (Appendix D, Figure 2).

Data Protection by Design and Impact Assessments For AI researchers using social media data, DPIAs are effectively mandatory under Article 35, as such research routinely triggers multiple high-risk criteria identified by the Article 29 Data Protection Working Party (2017): large-scale processing, dataset combination, and innovative technology deployment (Appendix B). A compliant DPIA must: (1) systematically describe the processing including data flows and infrastructure (Article 35(7)(a)); (2) assess necessity and proportionality while demonstrating Article 5 principle implementation (35(7)(b)); and (3) identify risks with appropriate mitigations (35(7)(c)). Notably, DPIAs are *living* documents requiring updates throughout iterative AI development cycles (Article 29 Data Protection Working Party 2017, p. 9).

Establishing a Legal Basis for Processing Social media researchers typically rely on either public interest (Article 6(1)(e)) or legitimate interests (Article 6(1)(f)) as their legal basis, with the choice determined by organisational context (Appendix D, Figure 3). Public authorities and universities operating under statutory research mandates can invoke public interest grounds, provided they demonstrate that their research serves a recognised public benefit (Court of Justice of the European Union 2024a). In contrast, private sector entities must conduct and document a *Legitimate Interest Assessment* (LIA) to justify their processing activities, as outlined in the Article 29 Data Protection Working Party (2017). Both require distinguishing broader research *interests* from specific processing *purposes*. The GDPR affords research significant flexibilities under Article 89 – including purpose compatibility, extended retention, and conditional rights limitations – provided researchers implement appropriate safeguards (European Data Protection Board 2021). These privileges enable research while maintaining privacy protection through proportionate technical and organisational measures.

IP and Contract Law IP and contract law determine the conditions under which social media data may be lawfully accessed, used and shared. Recent AI advances have intensified tensions in the IP landscape, raising questions about the scope of existing rights and exceptions (OECD 2025). Whilst scraping can implicate various rights (trademarks, trade secrets, publicity, moral rights), we focus on copyright and database rights.

Copyright and Database Rights Under the *InfoSoc Directive* (Directorate-General for Communications Networks, Content and Technology 2001), copyright grants authors exclusive economic rights over reproduction, distribution and public communication of their works. The CJEU has set a low originality threshold – 'the author's own intellectual creation' – holding in *Infopaq International A/S v Danske Dagblades Forening* (C-5/08) that even eleven-word extracts could qualify if bearing the author's personal stamp, and in *Painer v Standard Verlags GmbH* (C-145/10) that por-

trait photographs merit protection through creative choices in framing and lighting. This means even modest user-generated content (posts, short videos) may be protected if demonstrating individualised creative input. Simultaneously, the *Database Directive* (Council of the European Union 1996) establishes *sui generis* rights where 'substantial investment' exists in obtaining, verifying or presenting contents, potentially protecting platforms' aggregated collections of posts and metadata. This dual-layer protection – individual posts via copyright, aggregated collections via database rights – creates complex questions about scraping and reusing social media datasets, particularly where data provenance is unclear (OECD 2025). To balance these rights with scientific advancement, the *DSM Directive* (Council of the European Union 2019) introduced mandatory Text and Data Mining (TDM) exceptions: Article 3 grants research organisations unwaivable TDM rights for scientific research on lawfully accessed content, while Article 4 provides broader TDM rights that rightholders may reserve through machine-readable means such as the Robot Exclusion Protocol (`robots.txt`) – a text file that specifies which parts of a website automated crawlers may access.

Contractual and Technical Safeguards Contract law introduces an additional layer of legal obligations, often in the form of platform ToS or API agreements. These contractual terms define the conditions under which users, including researchers, may access and interact with platform data. Violating these terms can expose researchers to contractual liability, including account suspension, loss of data access, or even legal action for breach of contract. Whilst the DSM Directive renders certain contractual restrictions on research unenforceable, many limitations remain valid and binding. Platforms also implement `robots.txt` to control automated access, which researchers should respect as both a legal requirement (where they constitute Article 4 DSM reservations) and an ethical consideration.

2.3 Research Scope and Contributions

With these legal foundations established, this paper makes three core contributions. First, we introduce **PETLP**, a privacy-by-design framework that fundamentally reconceptualises ETL pipelines for the AI era. Rather than retrofitting compliance checkpoints, PETLP embeds DPIAs as living design tools that guide decisions from project conception through post-deployment. While no framework can accommodate every platform's unique policies and social dynamics, PETLP provides a methodological *template*. It offers general interpretive guidance rather than prescriptive instructions, empowering researchers to make informed decisions within their specific contexts.

Second, we provide operational clarity by synthesising three intersecting legal regimes – data protection, IP, and platform contract law – into practical decision trees (Appendix D, Figure 2-10). These trees are designed to resolve common ambiguities: when platform terms override statutory rights, how to qualify for DSM Article 3 protections, which privacy safeguards satisfy the GDPR, and whether models can be legally published.

Third, we demonstrate practical application through a systematic Reddit case study. We selected Reddit for its relatively open data access[2] and diverse community structure, which illuminates varied research ethics challenges. This implementation exemplar shows how PETLP navigates platform-specific complexities, providing a replicable model for platform-specific compliance analyses.

3 The PETLP Framework

3.1 Revisiting the ETL Model

Extract, Transform, Load (ETL) pipeline originated in the 1970s as a structured process for converting raw data into formats suitable for analysis and decision-making. At its core, ETL involves three sequential operations: *extracting* data from various sources, *transforming* it into a usable format, and *loading* it into a storage system for analysis. Given the absence of standardised practices for data collection and documentation in social media research (Bundtzen and Schwieter 2023), this model offers several attractive features: (1) providing a systematic structure for enhancing reproducibility and traceability; (2) well-established in both research and industry, facilitating methodological coherence; and (3) its sequential logic enabling a clear articulation of legal and ethical responsibilities at each stage.

However, applying ETL directly to social media research is not without difficulties. Research workflows are rarely linear. As Markham et al. (2012) observe, methodological stages often overlap or iterate unpredictably: researchers may extract and load data prior to transformation (*ELT*), delay processing due to ethical review cycles, or conduct exploratory analyses before finalising dataset structure.

Despite these limitations, we propose that an adapted ETL framework can serve as a valuable conceptual model for responsible data pipeline. We propose a modified framework – **PETLP** – that extends ETL in two key ways. First, the prefix "**P**" stands for *privacy-by-design*, implemented through the default use of DPIAs initiated prior to data collection and updated across all stages of the research lifecycle. This ex-ante approach enables researchers to identify and mitigate privacy risks before they materialise, rather than retrofitting safeguards after data collection. Notably, while PETLP provides structured checkpoints for privacy assessment, it accommodates the iterative nature of research by treating DPIAs as living documents that evolve with methodological refinements. Second, we introduce a fourth phase, *presentation*, supported by established research stage taxonomies (Markham et al. 2012) and social media research typologies (Bjerglund-Andersen and Söderqvist 2012) recognising dissemination as an indispensable component of the research lifecycle.

In the next sections, we examine each phase of the PETLP framework in detail:

- **Privacy-by-Design**: Operationalises GDPR Article 25 through default DPIAs as living documents that guide decisions throughout the research lifecycle.

[2]While X charges $200/month for 10,000 posts, Reddit permits 100 queries per minute under academic guidelines (Reddit, Inc. 2024b; X.com 2023).

- **Extract**: Analyses four data acquisition channels (platform-authorised, user-mediated, third-party, self-directed) with availability determined by researcher status, showing how copyright exceptions, platform terms, and GDPR obligations intersect differently for each method.
- **Transform**: Addresses the dual challenge of copyright reproductions (every preprocessing creates copies requiring authorisation) and privacy engineering (implementing minimisation, anonymisation attempts, and technical safeguards) during data cleaning and preparation.
- **Load**: Establishes secure storage architectures, international transfer compliance, and retention governance for transformed data, creating the infrastructure for compliant data access and querying.
- **Present**: Encompasses both AI model training and research dissemination (datasets, papers, models, services), evaluating extraction attack risks, copyright liability, platform restrictions, and the tension between open science mandates and privacy protection.

3.2 Privacy-by-Design

Privacy-by-design – mandated in Article 25 GDPR and reinforced by accountability obligations in Article 24 – requires privacy safeguards form part of a project's architecture from conception. PETLP operationalises this through DPIAs as *default design tools* spanning the entire research lifecycle. Researchers initiate DPIAs during pre-registration, update them when processing changes, and use them to guide decisions throughout every pipeline phase (Figure 1; see Appendix E for a comprehensive risk assessment example).

During pre-registration, researchers should consult comprehensive risk frameworks to identify potential privacy challenges early. Given the increasing adoption of LLMs in research contexts, Barberá (2025) offers particularly valuable guidance tailored to LLM privacy considerations: privacy and data protection risks with recommended mitigations (pp. 28–42), risk identification methodologies (pp. 48–56), estimation and evaluation protocols (pp. 57–65), control and mitigation strategies (pp. 66–73), residual risk assessment (p. 74), and state-of-the-art tools and benchmarks (as of March 2025) (pp. 97–101). Though LLM-focused, these frameworks provide transferable insights for broader AI applications, enabling researchers to anticipate risks during initial DPIA development rather than discovering them retrospectively.

This approach creates significant efficiencies, particularly for researchers relying on legitimate interests under Article 6(1)(f). As the UK Information Commissioner's Office clarifies, a comprehensive DPIA encompasses all elements of a Legitimate Interest Assessment (LIA) whilst providing deeper risk analysis (ICO 2023). The DPIA's structured evaluation of purpose, necessity, and proportionality automatically satisfies the LIA's three-part test (European Data Protection Board 2024a), eliminating redundant assessments while creating the documentary evidence required for ongoing compliance.

For social media research – which invariably triggers multiple high-risk indicators under EDPB criteria – this approach delivers five key benefits: (1) satisfies mandatory LIA requirements for legitimate interests processing; (2) fulfils Article 35 obligations for high-risk processing; (3) embeds risk-awareness and data minimisation directly into design decisions; (4) provides auditable compliance records for research projects; and (5) establishes structured mechanisms for refining safeguards as processing evolves.

By anticipating data sharing and replication requirements from inception, researchers can design access controls and anonymisation strategies that balance open science principles with privacy protection. This ensures research outputs satisfy both scholarly standards and regulatory requirements without requiring substantial retrospective modifications – converting compliance from perceived burden into methodological asset.

3.3 Extract

The *Extract* phase addresses the fundamental question of how researchers can lawfully acquire social media data for AI research. Unlike subsequent stages, extraction determines whether research may proceed at all. The legitimacy of data acquisition often serves as the gatekeeping criterion for entire research programmes.

Recent jurisprudence offers encouraging signals for academic researchers. The Hamburg District Court's decision in *LAION v. Kneschke* (2024) affirmed that creating `LAION-5B` dataset (Schuhmann et al. 2022) through web scraping constitutes legitimate Text and Data Mining (TDM) for scientific research purposes under the DSM Directive. Notably, the court held that platform terms cannot override statutory research exceptions for qualifying institutions – a principle codified in DSM Article 7(1).

However, this protection applies only to specific researchers under specific conditions. To navigate these variations, we develop a practice-informed typology of four extraction channels: platform-authorised access, user-mediated collection, third-party aggregation, and self-directed extraction. This typology, derived from regulatory analysis (what the law permits), platform architectures (what is technically feasible), and documented research practice (what researchers actually do), reveals that extraction legality depends on three intersecting factors: the researcher's institutional status (qualifying research organisation or commercial entity), the research purpose (scientific or commercial), and the specific legal basis invoked (DSM Article 3 and 4, and GDPR legal bases) (Appendix D, Figure 6).

Platform-authorised Access Platform-authorised access encompasses official APIs, research partnerships, and developer portals explicitly provided for sanctioned use. This method offers the clearest legal pathway, operating within platform-specific ToS and developer agreements. However, its availability and utility vary significantly by researcher type and platform strategy.

From a contractual perspective, platforms typically impose *browsewrap* agreements – terms that apply automatically upon site access. Their enforceability depends on ade-

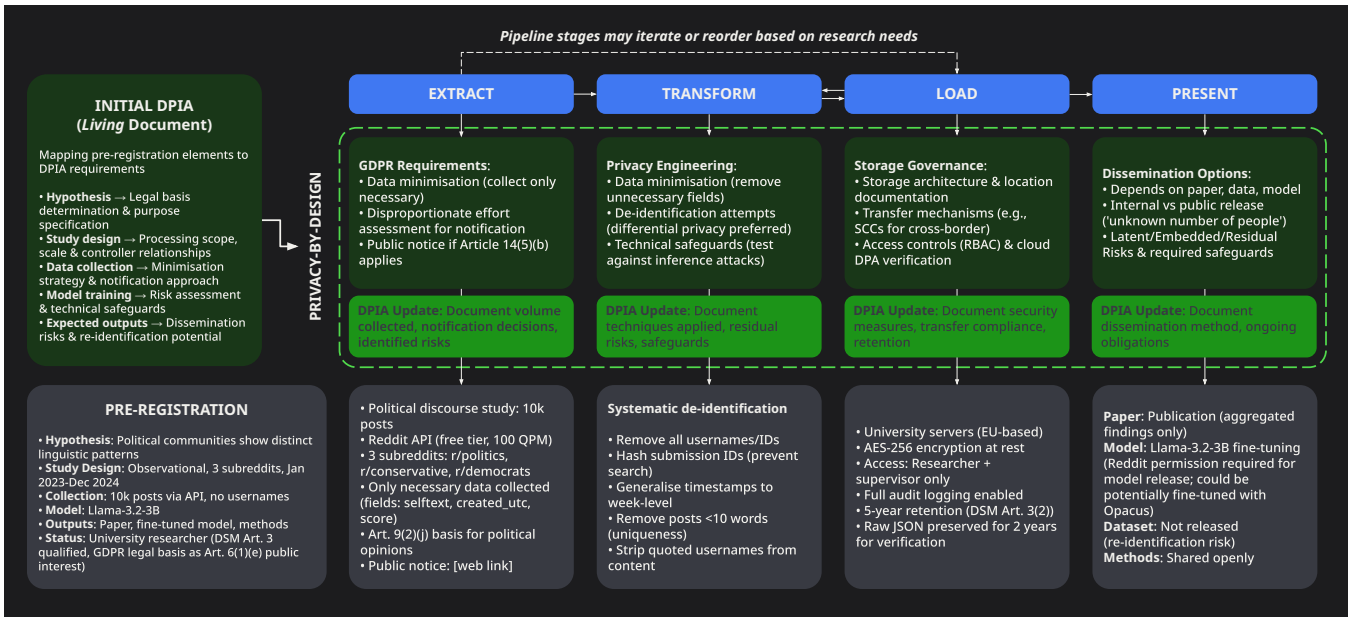

Figure 1: **Privacy-by-design ETLP (PETLP) framework for social media AI research**. *Pipeline stages (Appendix D, Figure 2-10) and Reddit case study (Appendix E).*

quate notice to users (Gupta 2012; Dou 2007). More robust are click-through agreements required for API access, which create explicit contractual relationships. Reddit exemplifies this tiered approach: whilst content remains publicly viewable, its browsewrap User Agreement prohibits unauthorised scraping (Section 7), effectively channelling researchers toward the official Data API, which requires explicit acceptance of additional Developer and API Terms (Reddit, Inc. 2024e,c,a).

For qualifying research organisations under DSM Article 3, these contractual restrictions may be legally unenforceable when conducting TDM for scientific research (Appendix D, Figure 4 and 5). The *LAION v. Kneschke* decision confirms that non-profit research institutions can invoke this mandatory exception. However, commercial researchers and those outside qualifying institutions remain bound by platform terms, as they must rely on the more limited Article 4 exception, which platforms can contractually exclude.

Despite its legal clarity, platform-authorised access is often ill-suited to the needs of academic research. Bruns (2021) identifies three structural limitations: (1) prohibitively expensive pricing models, (2) restricted export functionalities, and (3) market-oriented data schemas optimised for business analytics rather than open-ended scholarly inquiry. These limitations are not merely logistical; they exert substantive influence over what research questions can be asked and answered. For instance, Reddit's API imposes temporal restrictions that prevent longitudinal analysis beyond six months, fundamentally limiting research designs (Appendix I). In this way, platform-level design choices function as boundary conditions for academic inquiry, shaping not just the data available, but the entire epistemic scope of viable research.

User-mediated Data Collection User-mediated extraction methods rely on active participation by social media users who choose to contribute their data for research purposes. These include *data donations*, participation through browser extensions, or account-linked authorisations that enable researchers to retrieve user data under conditions of explicit consent. Under the GDPR Article 6(1)(a), such explicit consent can serve as a lawful basis for processing, provided it is freely given, specific, informed, and unambiguous. Such methods are often seen as ethically preferable because they foreground transparency, agency, and autonomy. Yet from a legal perspective, user consent does not necessarily override platform-imposed restrictions.

The well-documented case of NYU's *Ad Observer* project exemplifies the fragility of this method's legal status. Despite obtaining informed consent from users to collect and share advertising data, the project faced platform resistance. Meta shut down the researchers' accounts, citing violations of its ToS (Marshall 2021). This signals a regulatory ambiguity – wherein platforms retain contractual control over data flows, even in the face of user consent.

Pragmatically, this channel is also burdened by logistical and methodological obstacles. Recruiting a sufficiently large and demographically representative user base is often infeasible without major funding or institutional backing (Bundtzen 2023). Studies reliant on data donation or custom-built plug-ins tend to produce biased samples skewed toward technologically literate participants. This raises concerns about the generalisability of findings, particularly for research on platform-wide phenomena.

Third-party Aggregation Services Third-party data aggregators are companies or services that collect social media data on behalf of researchers, eliminating the need for

researchers to scrape or access platforms directly. These services – ranging from commercial vendors like Bright Data to research-oriented platforms like Pushshift (now defunct) – gather posts, comments, and metadata from social media sites, then provide this pre-collected data to researchers either for free or for a fee. While appealing for their convenience and scalability, these services occupy a precarious legal position.

There are no widely reported cases in the EU that are directly equivalent to *Meta v. Bright Data*, where a court ruled on the third party legality of web scraping public data from social media platforms. However, the EU landscape proves more restrictive. Under the Database Directive Article 7, platforms hold rights over substantial investments in data organisation, prohibiting extraction of substantial parts without authorisation. Article 7(5) extends this to 'repeated and systematic extraction' of even insubstantial parts that conflict with normal database exploitation.

Beyond jurisdictional differences, these services face contractual and data protection challenges. Reddit's user agreement, whilst granting the platform broad content rights, explicitly prohibits third-party redistribution without permission (Reddit, Inc. 2024e). The GDPR further complicates matters, requiring aggregators to establish a lawful basis under Article 6 and implement heightened safeguards for special category data under Article 9.

This suggests that datasets on platforms like Academic Torrents, despite their widespread academic use, exemplify these converging legal risks. Following Reddit's enforcement action against Pushshift, vast archives of Reddit data now circulate via torrents, typically without licenses beyond warnings that content 'may be protected by copyright' (Academic Torrent 2025). This situation presents researchers with an ethical paradox: datasets fundamental to research exist in legal limbo, potentially violating platform terms, database rights, and data protection law. That these datasets migrated from Pushshift to torrents following platform enforcement underscores their legally dubious status.

Self-directed Extraction Self-directed extraction – where researchers control their own scraping or automation – provides the greatest flexibility in data collection but also carries the highest legal risk. Academic and commercial researchers face fundamentally different regulatory landscapes when using this method.

For qualifying research organisations, DSM Article 3 provides robust protection. The *LAION v. Kneschke* court explicitly held that extracting publicly available image URLs for AI dataset constitutes legitimate TDM for research. Platforms may implement measures to ensure the security and integrity of the networks and databases, but these 'shall not go beyond what is necessary' (Article 3(3)). Rate limiting might be proportionate; blanket automated access bans likely exceed necessity.

Commercial researchers face a sharply different landscape. Without Article 3 protection, they must rely on Article 4's general TDM exception, which platforms can exclude through machine-readable opt-outs. Reddit's `robots.txt` and ToS constitute such opt-outs, rendering commercial scraping legally impermissible without prior approval.

Meanwhile, all researchers – whether academic or commercial – must recognise that the GDPR operates independently of copyright exceptions. The *LAION v. Kneschke* decision of TDM rights addresses only intellectual property; data protection obligations remain fully applicable. As established in our default DPIA approach, researchers must demonstrate lawful basis, implement data minimisation, and maintain comprehensive documentation throughout the pipeline. The EDPB specifically flags web scraping's risks – large volume of data collected, the large number of data subjects, and the indiscriminate collection (European Data Protection Board 2024b, Paragraph 86) – requiring heightened safeguards regardless of TDM authorisation.

3.4 Transform

The *Transform* phase focuses on converting raw social media data into analysis-ready formats through cleaning, validation, normalisation, and restructuring operations. Unlike extraction, transformation engages specific legal challenges around intermediate data processing. Each cleaning operation – removing duplicates, standardising formats, extracting features, or restructuring databases – creates new copyright reproductions whilst simultaneously constituting the GDPR processing activities. This phase requires researchers to implement DPIA-identified safeguards *and* ensuring their preprocessing activities remain within applicable IP exceptions (Appendix D, Figure 7).

Data Preprocessing and IP rights Data transformation necessarily creates multiple reproductions falling within InfoSoc Article 2's broad scope. Every preprocessing step – loading data into memory for cleaning, creating intermediate files during normalisation – constitutes reproduction requiring authorisation. While Article 5(1) InfoSoc exempts 'transient or incidental' copies forming an integral and essential part of a technological process, cleaned datasets and extracted features have independent value beyond mere technical facilitation.

For qualifying research organisations, DSM Article 3 permits these reproductions when conducting TDM for scientific research. The *LAION v. Kneschke* decision (Hamburg District Court, 2024) confirms this covers dataset preparation activities, with the court ruling that downloading, analysing, and restructuring data for research purposes falls within the TDM exception. Article 3(2) explicitly protects intermediate datasets, allowing retention of processed copies 'for the purposes of scientific research, including for the verification of research results'.

Commercial researchers, however, face significant constraints. Without Article 3's mandatory protection, they must navigate Article 4's opt-out provisions. Reddit exemplifies comprehensive opt-out strategies, combining technical measures (`robots.txt: User-agent: *; Disallow: /`), contractual prohibitions (Reddit, Inc. 2024e, Section 7), and API restrictions (Reddit, Inc. 2024e). The evolution of "machine-readable" opt-outs further complicates matters. While the *LAION v. Kneschke* court noted *in obiter* (a non-binding observation that could nonetheless influence fu-

ture interpretations) that natural language statements might qualify as machine-readable, initiatives like `llms.txt` (Howard 2024) illustrate how machine-readability might evolve beyond traditional interpretations to include human-readable but parser-friendly formats.

Privacy-Preserving Transformation The Transform phase operationalises GDPR principles into concrete preprocessing decisions. Unlike raw data extraction, transformation offers opportunities to embed privacy safeguards. Each decision – which fields to retain, how to aggregate data, whether to infer missing values – must balance research utility against privacy principles.

Article 5(1)(c) requires data minimisation: retaining only data 'adequate, relevant and limited to what is necessary'. For transformation, this means actively removing unnecessary fields, aggregating where individual-level detail is not required, and resisting the temptation to preserve potentially useful data without clear purpose.

Transformation also provides the primary opportunity for anonymisation – though achieving true anonymity proves challenging. Removing usernames and IDs (pseudonymisation) offers minimal protection; high-dimensional social media data remains highly re-identifiable. De Montjoye et al. (2013) demonstrated that just four spatial-temporal data points were enough to uniquely identify 95% of individuals in a 1.5 million-person mobility dataset. Techniques like k-anonymity, once standard, have been shown to suffer from both attribute and linkage vulnerabilities, particularly at scale (Sweeney 2002; Gadotti et al. 2024).

Against this backdrop, *differential privacy* (DP) has emerged as the state-of-the-art approach for anonymisation, offering provable guarantees against a broad range of adversarial attacks (Jiang et al. 2021). By adding calibrated noise to aggregated statistics or creating synthetic datasets, DP enables privacy-preserving transformation. Real-world implementations include Social Science One's DP-protected link-sharing statistics on Facebook (Nayak 2020), Wikipedia's pageview metrics (Adeleye et al. 2023), and Apple's DP-enabled word detection from Reddit comment histories (Hu et al. 2023). However, researchers must recognise that even summary-level data or aggregated features can leak private information. Membership inference, model inversion, and linkage attacks have demonstrated that "aggregate" does not equal "anonymous" (Gadotti et al. 2024). These vulnerabilities underscore why transformation requires careful consideration beyond simple technical implementation.

Indeed, for research under Article 89, transformation represents the key phase for implementing 'appropriate safeguards'. This encompasses both technical measures (encryption, access controls, differential privacy) and *organisational* ones – documenting transformation logic, maintaining processing logs, and ensuring reproducibility without compromising privacy. Researchers should view transformation not as routine data cleaning but as a proactive opportunity to embed privacy-by-design principles directly into research datasets.

3.5 Load

The *Load* phase migrates transformed social media data into secure storage systems – databases or cloud repositories – where it becomes accessible for querying and analysis. Unlike the Presentation phase where model training occurs, Load focuses exclusively on establishing compliant storage infrastructure and *internal governance*. Here, the safeguards defined in the Transform stage are encoded into concrete storage architectures, access protocols, and retention schedules (Appendix D, Figure 8).

Data Accessibility and Transfer Transformed data requires responsible access controls, consistent with the GDPR principles. Under the GDPR, *fully anonymised data* (Recital 26) falls outside its scope. However, the EDPB warns that achieving true anonymisation with social media data proves difficult (Article 29 Data Protection Working Party 2017), meaning most stored datasets remain *pseudonymised* and subject to transfer restrictions under Articles 45–49 GDPR: (1) *Adequacy decisions* (Article 45) allow transfers to countries deemed by the European Commission to offer 'essentially equivalent' protection to require no additional safeguards; (2) *Appropriate safeguards* (Article 46) apply when transfers to non-adequate countries (e.g., the US, unless covered by the EU-US Data Privacy Framework) require Standard Contractual Clauses (SCCs), Binding Corporate Rules (BCRs), or administrative arrangements; and (3) *Derogations* (Article 49) permit, in exceptional circumstances, explicit consent or public interest justifications to be invoked – but only for non-repetitive, non-systematic transfers.

These obligations apply when loading data into storage systems across borders – common in distributed research collaborations. The DPIA must document transfer mechanisms as fundamental steps to demonstrate that personal data is not processed unlawfully (European Data Protection Board 2024b, Paragraph 56), particularly when using cloud storage providers with global infrastructure.

Storage Architecture and Security Implementation Security obligations arise immediately upon data storage. Loading data into persistent storage triggers Article 32's mandate for security measures 'appropriate to the risk'. The EDPB emphasises that storage systems must implement state-of-the-art protections regardless of complexity or cost (European Data Protection Board 2024c, Paragraph 7). Essential measures include, but not limited to, role-based access controls limiting data to authorised researchers, encryption both at rest (e.g., Advanced Encryption Standard AES-256) and in transit (e.g., Transport Layer Security TLS 1.3).

Cloud storage requires additional contractual safeguards. Major cloud providers offer pre-approved Data Processing Addenda meeting Article 28 requirements (Amazon Web Services, Inc. 2023; Google LLC 2025; Microsoft Corporation 2025; Supabase, Inc. 2025). When loading data into these platforms, researchers should verify agreements cover their specific storage architecture.

Retention Scheduling and Deletion Protocols Loading data into persistent storage triggers the storage limitation

principle under Article 5(1)(e) GDPR, which requires personal data be kept no longer than necessary for the specified purposes. Whilst the GDPR permits extended retention for scientific research under Article 89(1) safeguards, *and* DSM Article 3(2) allows qualifying research organisations to retain data 'for the purposes of scientific research, including for the verification of research results', these provisions require operationalisation through concrete retention policies.

Researchers should implement: (1) defined retention schedules specifying maximum storage periods per data category (e.g., 5 years for processed datasets, 2 years for raw API responses); (2) automated deletion triggers using database management tools to enforce these schedules; (3) audit trails documenting retention decisions and deletions; and (4) exception handling procedures for data subject to legal holds or verification requirements.

3.6 Present

The *Present* stage marks the transition from internal governance to public dissemination. It is the point at which insights, outputs, or models derived from social media data are shared with external audiences – whether through academic publications, public datasets, or open-source distribution of trained models. Whilst earlier stages of the PETLP framework emphasised data acquisition and internal processing controls, this final phase introduces distinct legal and ethical challenges associated with *external exposure*.

The risks introduced at this stage cannot be retroactively addressed by upstream compliance efforts alone. This section examines how researchers must navigate the competing demands of privacy protection when sharing findings, open science mandates requiring data accessibility, and copyright restrictions limiting redistribution – challenges that emerge uniquely at the point of public dissemination (Appendix D, Figure 9 and 10).

Data Distribution Disseminating research findings requires balancing scientific transparency against re-identification risks and platform redistribution restrictions.

Re-identification Risks Disseminating research findings requires careful alignment with the GDPR's core principles – especially purpose limitation (Article 5(1)(b)) and data protection by design (Article 25). The Present stage focuses on how outputs may indirectly re-expose personal data, including through seemingly benign presentation methods.

For data distribution, a primary concern is the citation of verbatim content from social media posts. Adams (2022) demonstrates that such quotations – even when stripped of usernames – enable re-identification through search engines. This risk is especially pronounced in niche subreddits or sensitive thematic contexts, where content is technically public but effectively traceable.

To mitigate this, researchers should calibrate their dissemination methods based on the content's identifiability and subject sensitivity. Strategies may include *paraphrasing* user content instead of quoting verbatim, *aggregation* of findings to group-level insights, *synthetic illustration* using fabricated yet plausible data to demonstrate patterns, and *visualisation methods* that omit individual-level markers.

These decisions must be pre-embedded in the project's DPIA and justified under Article 89's 'appropriate safeguards'. Documentation should explicitly address how publication methods prevent the re-identification risks and surveillance effects identified by the EDPB.

Transparency-Restriction Tensions Public presentation requires navigating tensions between transparency mandates and platform restrictions. Reddit's Developer Terms prohibit third-party data redistribution (Reddit, Inc. 2024c, Section 7.4), whilst major funders like Horizon Europe mandate open access to underlying data (European Innovation Council and SMEs Executive Agency 2024).

Although the *LAION v. Kneschke* decision affirmed that TDM copyright exceptions cover dataset *creation*, it did not extend this protection to *distribution*. The court's narrow focus on reproduction rights leaves researchers without clear legal grounds to redistribute extracted content, even when initial extraction was lawful under DSM Article 3.

For researchers, practical approaches to balance openness with compliance include: (1) *hydration via post IDs*, publishing only content identifiers, allowing reconstruction within platform policies (e.g., RetweetBERT (Jiang, Ren, and Ferrara 2023)); (2) *synthetic datasets*, creating statistically equivalent data without redistributing actual posts (e.g., SynthPAI (Yukhymenko et al. 2024)); (3) *secure analysis environments*, providing controlled remote access rather than data downloads (e.g., SANE (SURF n.d.)); and (4) *platform programmes*, leveraging official research access channels (e.g., `reddit4researchers` (Reddit, Inc. 2024d)).

Researchers should document their dissemination strategy within the DPIA, specifying how it balances legal compliance, data minimisation, and research transparency.

AI model Distribution Publishing AI models introduces persistent risks as models may retain personal information, reproduce copyrighted content, and violate platform contractual obligations.

Embedded Privacy Risks The EDPB warns that 'AI models trained on personal data cannot, in all cases, be considered anonymous' (European Data Protection Board 2024b, Paragraph 34), meaning that model publication may constitute ongoing personal data processing. This concern is reinforced by the observation that 'information from the training dataset, including personal data, may still remain *absorbed* in the parameters of the model' (European Data Protection Board 2024b, Paragraph 31). Together, these warnings highlight that model release carries inherent privacy risks that persist beyond the training phase.

While differential privacy techniques such as Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al. 2016), implemented through tools like `Opacus` (Aketi et al. 2025), offer theoretical protection against membership inference, their practical deployment varies significantly by model scale. For smaller models ($<$100M parameters) typical of classification tasks, `Opacus` provides effective privacy protection with minimal utility loss. Medium-scale models like BERT (100M-1B parameters) maintain acceptable performance, particularly when fine-tuning rather than training from scratch. Recent advances including FSDP sup-

port, and LoRA for parameter-efficient fine-tuning now enable scaling to LLMs like `Llama-3-8b`. However, Aketi et al. (2025) focus solely on demonstrating feasibility and computational efficiency for large models, leaving the critical question of model quality and utility under privacy constraints unexplored.

Copyright Liability Model publication introduces copyright risks beyond privacy concerns. We re-emphasise that the *LAION v. Kneschke* decision distinguished dataset creation from model training, explicitly declining to rule whether training and distributing AI models falls within TDM exceptions. This judicial restraint leaves researchers uncertain about model publication rights, even when dataset creation was lawful. Until the pending CJEU case (*Like Company v. Google Ireland*, C-250/25) provides clarity, researchers face three liability scenarios (Rosati 2024): (1) *direct liability* when published models generate outputs substantially reproducing training content; (2) *secondary liability* when third parties use published models to create infringing content; and (3) *distribution liability* for sharing datasets containing copyrighted material. Importantly, contractual disclaimers accompanying research outputs provide limited protection. While researchers might attempt to exclude their liability through such provisions (e.g., standardised open-source licences to specify liability limitations), these contractual arrangements typically cannot override statutory copyright protections (Rosati 2024). Legal liability may persist despite explicit disclaimers.

Contractual Constraints For social media research, platform-specific restrictions compound these risks. Reddit prohibits using its data for AI training without explicit permission (Reddit, Inc. 2024c, Section 4.2). Although qualifying research organisations may invoke DSM Article 3 for dataset creation, the uncertainty around model training and publication means researchers should implement technical measures to prevent models from reproducing training data verbatim, or consider publishing model architectures and training procedures rather than trained weights.

4 Discussion

While PETLP successfully navigates overlapping legal regimes as validated through our Reddit case study, it cannot resolve the structural tensions inherent in social media AI research governance.

From Checkbox to Continuous Compliance: Traditional AI governance artifacts – Model Cards, Data Cards, Transparency Notes – often deliver only "checkbox compliance", satisfying regulatory requirements while failing to enable meaningful oversight (Kawakami, Wilkinson, and Chouldechova 2024). PETLP's living DPIA approach embeds compliance decisions directly into research workflows, shifting from performative to operational transparency. This treats privacy as a design principle shaping methodological choices from extraction through deployment. However, even this approach operates within an ecosystem where transparency often legitimises rather than scrutinises AI systems.

Navigating Fragmented Governance: The regulatory landscape lacks coherence, with governance instruments emerging daily from disparate actors (Arnold et al. 2024). PETLP structures this complexity through decision trees, yet cannot eliminate underlying fragmentation. The `LAION-5B` case illustrates a related challenge. Even when datasets satisfy one legal domain (DSM Article 3), they may violate others (the dataset's deprecation following discovery of illegal content (Thiel, Stroebel, and Portnoff 2023; NeurIPS 2025)). Researchers must therefore navigate not just fragmented governance but also the gaps between legal domains – where copyright compliance, platform terms, content regulations, and data protection operate independently.

Platform Heterogeneity as Structural Barrier: While *LAION v. Kneschke* protects academic scraping legally, technical enforcement varies unpredictably. Reddit's six-month API limitation constrains longitudinal research; X's pricing excludes unfunded researchers; Meta's account suspensions create chilling effects. PETLP acknowledges this through four extraction channels with distinct trade-offs, yet each platform demands bespoke implementation, multiplying complexity for multi-platform studies.

Operationalising Responsible Practice: PETLP requires complementary infrastructure for practical deployment. First, institutional support through template DPIAs and integrated legal-ethical review would reduce researcher burden. Universities could extend IRB processes to encompass DPIA assessment. Second, shared computational environments with pre-configured differential privacy would enable wider adoption of privacy protections, addressing gaps between theoretical safeguards and practical deployment given documented PII leakage (Nasr et al. 2023). While infrastructures like European Open Science Cloud (EOSC) demonstrate federated research platforms, they currently lack privacy tooling. Third, automated compliance tools could bridge legal complexity and research practice. Provenance tracking systems that generate GDPR-compliant documentation while qualifying for regulatory safe harbours (Longpre et al. 2024) would transform PETLP's decision trees into executable compliance pipelines. A forthcoming `RedditHarbor` toolkit demonstrates one path forward – automating PETLP's compliance logic specifically for Reddit research. These developments recognise that responsible AI research requires not just frameworks but ecosystems where institutional support, technical infrastructure, and automated tools converge to make compliance achievable rather than aspirational.

5 Conclusion

PETLP repositions privacy and compliance from constraints to design principles that enhance research rigour and public trust. Through detailed legal analysis and practical implementation, we demonstrate that responsible research practice strengthens rather than compromises scientific innovation – creating conditions for sustainable, trustworthy AI development that withstands regulatory scrutiny and maintains social licence. We position PETLP not as a definitive solution but as a structured foundation for ongoing dialogue about how AI research can serve societal benefit whilst respecting individual privacy in an era where these values increasingly conflict.

References

2007. DOUGLAS v. Talk America Inc., a Pennsylvania corporation, Real Party in Interest. Decided on July 18, 2007.

2010. Romano v. Steelcase Inc. 2010 NY Slip Op 20388, 30 Misc 3d 426, published by New York State Law Reporting Bureau pursuant to Judiciary Law § 431.

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Academic Torrent. 2025. Reddit subreddits metadata, rules and wikis 2025-01.

Adams, N. N. 2022. 'Scraping' Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of COVID-19. *International Journal of Social Research Methodology*.

Adeleye, T.; Berghel, S.; Desfontaines, D.; Hay, M.; Johnson, I.; Lemoisson, C.; Machanavajjhala, A.; Magerlein, T.; Modena, G.; Pujol, D.; et al. 2023. Publishing Wikipedia usage data with strong privacy guarantees. *arXiv preprint arXiv:2308.16298*.

Aketi, S. A.; Bullock, W.; Kalemaj, I.; Ullah, E.; and Zhang, H. 2025. Scaling Private Deep Learning with Opacus: Advances for Large Language Models. In *Championing Open-source DEvelopment in ML Workshop @ ICML25*.

Alhamed, F.; Ive, J.; and Specia, L. 2024. Using large language models (LLMs) to extract evidence from pre-annotated social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 232–237.

Amazon Web Services, Inc. 2023. AWS Data Processing Addendum (DPA) - Navigating GDPR Compliance.

Ananthakrishnan, G.; Jayaraman, A. K.; Trueman, T. E.; Mitra, S.; AK, A.; and Murugappan, A. 2022. Suicidal intention detection in tweets using BERT-based transformers. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 322–327. IEEE.

Arnold, Z.; Schiff, D. S.; Schiff, K. J.; Love, B.; Melot, J.; Singh, N.; Jenkins, L.; Lin, A.; Pilz, K.; Enweareazu, O.; et al. 2024. Introducing the AI Governance and Regulatory Archive (AGORA): An Analytic Infrastructure for Navigating the Emerging AI Governance Landscape. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 39–48.

Article 29 Data Protection Working Party. 2017. Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679.

Barberá, I. 2025. AI Privacy Risks & Mitigations – Large Language Models (LLMs). Technical report, Support Pool of Experts Programme, European Data Protection Board (EDPB).

Batrinca, B.; and Treleaven, P. C. 2015. Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30: 89–116.

Bjerglund-Andersen, N.; and Söderqvist, T. 2012. Social Media and Public Health Research. Technical report, University of Copenhagen.

Boyd, D. M.; and Ellison, N. B. 2007. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1): 210–230.

Brown, M. A.; Gruen, A.; Maldoff, G.; Messing, S.; Sanderson, Z.; and Zimmer, M. 2024. Web scraping for research: Legal, ethical, institutional, and scientific considerations. *arXiv preprint arXiv:2410.23432*.

Bruns, A. 2021. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Disinformation and Data Lockdown on Social Platforms*, 14–36.

Bundtzen, S. 2023. Data Access.

Bundtzen, S.; and Schwieter, C. 2023. Access to Social Media Data for Public Interest Research: Lessons Learnt & Recommendations for Strengthening Initiatives in the EU and Beyond. Technical report, Institute for Strategic Dialogue (ISD Germany).

Chiauzzi, E.; and Wicks, P. 2019. Digital trespass: ethical and terms-of-use violations by researchers accessing data from an online patient community.

Council of the European Union. 1996. Directive 96/9 on the Legal Protection of Databases.

Council of the European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market. [2019] OJ L 130/92.

Court of Justice of the European Union. 2022. Case C-184/20, OT v Vyriausioji tarnybinės etikos komisija. ECLI:EU:C:2022:601, Judgment of the Court (Grand Chamber) of 1 August 2022, Request for a preliminary ruling from the Vilniaus apygardos administracinis teismas.

Court of Justice of the European Union. 2023a. Case C-21/23, Lindenapotheke: Request for a preliminary ruling from the Bundesgerichtshof (Germany) lodged on 19 January 2023 — ND v DR. OJ C 155, 2.5.2023.

Court of Justice of the European Union. 2023b. Case C-252/21, Meta Platforms Inc and Others v Bundeskartellamt. ECLI:EU:C:2023:537, Judgment of the Court (Grand Chamber) of 4 July 2023, Request for a preliminary ruling from the Oberlandesgericht Düsseldorf.

Court of Justice of the European Union. 2024a. Case C-200/23, *Agentsia po vpisvaniyata* v OL. ECLI:EU:C:2024:827, Judgment of 4 October 2024.

Court of Justice of the European Union. 2024b. Case C-446/21, Maximilian Schrems v. Meta Platforms Ireland Limited. ECLI:EU:C:2024:834, Judgment of 4 October 2024.

De Montjoye, Y.-A.; Hidalgo, C. A.; Verleysen, M.; and Blondel, V. D. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1): 1376.

Deng, X.; Bashlovkina, V.; Han, F.; Baumgartner, S.; and Bendersky, M. 2023. LLMs to the moon? Reddit market sentiment analysis with large language models. In *Companion Proceedings of the ACM Web Conference 2023*, 1014–1019.

Directorate-General for Communications Networks, Content and Technology. 2001. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. [2001] OJ L 167/10, Article 5(1).

Dornis, T. W.; and Stober, S. 2025. Generative AI Training and Copyright Law. *arXiv preprint arXiv:2502.15858*.

European Commission. 2021. Ethics and Data Protection. Published: 05 July 2021.

European Data Protection Board. 2020. Guidelines 07/2020 on the Concepts of Controller and Processor in the GDPR.

European Data Protection Board. 2021. Study on the Secondary Use of Personal Data in the Context of Scientific Research. Published by the EDPB.

European Data Protection Board. 2024a. Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, Version 1.0. Adopted on 8 October 2024.

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.

European Data Protection Board. 2024c. Report of the work undertaken by the ChatGPT Taskforce.

European Data Protection Supervisor. 2021. Guidelines 8/2020 on the targeting of social media users. Published 7 July 2021.

European Digital Media Observatory. 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. Technical report.

European Innovation Council and SMEs Executive Agency. 2024. Your guide to open science in Horizon Europe.

Gadotti, A.; Rocher, L.; Houssiau, F.; Creţu, A.-M.; and De Montjoye, Y.-A. 2024. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29): eadn7053.

Garcia-Pueyo, L.; Kumar Sunkara, V.; Senthil Kumar, P.; Diwan, M.; Ge, Q.; Javaherian, B.; and Verroios, V. 2023. Detecting and Limiting Negative User Experiences in Social Media Platforms. In *Proceedings of the ACM Web Conference 2023*, 4086–4094.

Google LLC. 2025. Cloud Data Processing Addendum.

Gupta, I. 2012. Are websites adequately communicating terms & conditions link in a browse-wrap agreement? *European Journal of Law and Technology*, 3(2).

He, L.; Braggaar, A.; Basar, E.; Krahmer, E.; Antheunis, M.; and Wiers, R. 2024. Exploring user engagement through an interaction lens: what textual cues can tell us about human-chatbot interactions. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 1–14.

Howard, J. 2024. llms.txt: The /llms.txt file - a proposal to standardise using an /llms.txt file to inform Large Language Models for better website understanding.

Hu, L.; Habernal, I.; Shen, L.; and Wang, D. 2023. Differentially private natural language models: Recent advances and future directions. *arXiv preprint arXiv:2301.09112*.

ICO. 2023. How do we apply legitimate interests in practice?

Iqbal, H.; Khan, U. M.; Khan, H. A.; and Shahzad, M. 2022. Left or right: A peek into the political biases in email spam filtering algorithms during us election 2020. In *Proceedings of the ACM Web Conference 2022*, 2491–2500.

Ji, S.; Pan, S.; Li, X.; Cambria, E.; Long, G.; and Huang, Z. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1): 214–226.

Jiang, H.; Pei, J.; Yu, D.; Yu, J.; Gong, B.; and Cheng, X. 2021. Applications of differential privacy in social network analysis: A survey. *IEEE transactions on knowledge and data engineering*, 35(1): 108–127.

Jiang, H.; Yu, J.; Hu, C.; Zhang, C.; and Cheng, X. 2018. SA framework based de-anonymization of social networks. *Procedia Computer Science*, 129: 358–363.

Jiang, J.; Ren, X.; and Ferrara, E. 2023. Retweet-bert: political leaning detection using language features and information diffusion on social networks. In *Proceedings of the international AAAI conference on web and social media*, volume 17, 459–469.

Kawakami, A.; Wilkinson, D.; and Chouldechova, A. 2024. Do Responsible AI Artifacts Advance Stakeholder Goals? Four Key Barriers Perceived by Legal and Civil Stakeholders. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 670–682.

Longpre, S.; Mahari, R.; Obeng-Marnu, N.; Brannon, W.; South, T.; Gero, K. I.; Pentland, A.; and Kabbara, J. 2024. Position: Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them? In *Forty-first International Conference on Machine Learning*.

Lynskey, O. 2015. *The foundations of EU data protection law*. Oxford University Press.

Markham, A.; Buchanan, E.; with feedback from the AOIR Ethics Working Committee; et al. 2012. Ethical decision-making and internet research: Recommendations from the AoIR Ethics Working Committee (Version 2.0).

Marshall, E. 2021. Why Facebook's Claims About the Ad Observer Are Wrong.

Microsoft Corporation. 2025. Microsoft Products and Services Data Protection Addendum (DPA).

Morten, C. J.; Nicholas, G.; and Viljoen, S. 2024. Researcher access to social media data: Lessons from clinical trial data sharing. *Berkeley Tech. LJ*, 39: 109.

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*.

Nayak, C. 2020. New privacy-protected Facebook data for independent research on social media's impact on democracy. *Facebook Research*.

Nesterov, A.; Hollink, L.; and van Ossenbruggen, J. 2024. How contentious terms about people and cultures are used

in linked open data. In *Proceedings of the ACM on Web Conference 2024*, 4523–4533.

NeurIPS. 2025. Deprecated Datasets. https://neurips.cc/public/deprecated-datasets.

Nguyen, T.-P.; Razniewski, S.; Varde, A.; and Weikum, G. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, 1907–1917.

Obar, J. A.; and Wildman, S. 2015. Social media definition and the governance challenge: An introduction to the special issue.

OECD. 2025. Intellectual property issues in artificial intelligence trained on scraped data. Technical Report 33.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2: 13.

Qian, J.; Li, X.-Y.; Jung, T.; Fan, Y.; Wang, Y.; and Tang, S. 2019. Social network de-anonymization: More adversarial knowledge, more users re-identified? *ACM Transactions on Internet Technology (TOIT)*, 19(3): 1–22.

Reddit, Inc. 2024a. Data API Terms - Reddit.

Reddit, Inc. 2024b. Reddit Data API Wiki. Accessed: 2024-09-12.

Reddit, Inc. 2024c. Reddit Developer Terms. Accessed: 2024-09-20.

Reddit, Inc. 2024d. Reddit for Researchers Beta Program: We're Live!

Reddit, Inc. 2024e. Reddit User Agreement. Accessed: 2024-09-17.

Rosati, E. 2024. Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law. *European Journal of Risk Regulation*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Solovev, K.; and Pröllochs, N. 2022. Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In *Proceedings of the ACM Web Conference 2022*, 3656–3661.

Stieglitz, S.; Mirbabaie, M.; Ross, B.; and Neuberger, C. 2018. Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39: 156–168.

Supabase, Inc. 2025. Data Processing Addendum (DPA).

SURF. n.d. SANE: Secure Environment for Analysing Sensitive Data. Accessed: April 19, 2025.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05): 557–570.

Taylor, J.; and Pagliari, C. 2018. Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, 14(2): 1–39.

Terzis, P.; Veale, M.; and Gaumann, N. 2024. Law and the Emerging Political Economy of Algorithmic Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1255–1267.

Thiel, D.; Stroebel, M.; and Portnoff, R. 2023. Generative ML and CSAM: Implications and mitigations. *Thorn & Stanford Internet Observatory*.

Tian, W.; Mao, J.; Jiang, J.; He, Z.; Zhou, Z.; and Liu, J. 2018. Deeply understanding structure-based social network de-anonymization. *Procedia Computer Science*, 129: 52–58.

Vuruma, S. K. R.; Wu, D.; Sen Gupta, S.; Aust, L.; Lookingbill, V.; Henry, C.; Ren, Y.; et al. 2024. Utilizing Large Language Models to Identify Reddit Users Considering Vaping Cessation for Digital Interventions. *arXiv preprint arXiv:2404.17607*.

X.com. 2023. Rate Limits. Accessed: 2024-09-14.

Xu, P. 2024. Tight Competitive and Variance Analyses of Matching Policies in Gig Platforms. In *Proceedings of the ACM on Web Conference 2024*, 5–13.

Yukhymenko, H.; Staab, R.; Vero, M.; and Vechev, M. 2024. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37: 120735–120779.