Natural Language Processing (2026), **32**, pp. 662–675
doi:10.1017/nlp.2026.10036

CAMBRIDGE UNIVERSITY PRESS

RESPONSIBLE NLP

The same model, different responsibilities: Responsible NLP under the EU AI Act

Salima Lamsiyah

University of Luxembourg, Luxembourg
Email: salima.lamsiyah@uni.lu

(Received 2 August 2026 UTC; revised 2 August 2026 UTC; accepted 6 August 2026 UTC;
first published online 22 September 2026 UTC)

Abstract

The European Union's Artificial Intelligence Act changes the governance environment in which natural language processing (NLP) systems are developed and deployed. Its central lesson for Responsible NLP is that responsibility does not follow automatically from a model's architecture or benchmark score. It is distributed across three layers: a general-purpose model, the task-specific system built around it, and the institution in which that system acts. This piece analyses those layers through recruitment, education, public-service chatbots, public-interest text generation, and general-purpose language models. It shows how the Act turns principles into duties concerning risk management, data governance, documentation, transparency, human oversight, and robustness, together with complaint, explanation, and procedural-accountability mechanisms, while also examining its interaction with data-protection law. The analysis reveals a deeper research challenge: compliance requires an evidence chain connecting intended purpose, affected people, evaluation, institutional workflow, and post-deployment monitoring. Regulation cannot create that evidence by itself. Multilingual fairness, construct validity, meaningful explanation, uncertainty communication, and effective contestability remain scientific and sociotechnical problems. The EU AI Act should therefore be understood as a legal floor for Responsible NLP, not its scientific ceiling.

Keywords: Responsible NLP; EU AI Act; transparency; high-risk AI; human oversight

Legal-status note. References to the AI Act in this piece mean Regulation (EU) 2024/1689 as amended by Regulation (EU) 2026/1744. The legal and implementation materials are current to 1 August 2026 (European Parliament and Council of the European Union 2024, 2026).

1. One model, several legal lives

Imagine the same language model in three settings. In the first, it summarises meeting notes. In the second, it ranks job applications. In the third, it helps a public authority decide whether a person qualifies for social assistance. The underlying model may be identical. The legal and ethical responsibilities are not.

This is the most useful starting point for understanding the EU AI Act from an NLP perspective. Regulation (EU) 2024/1689, as amended by Regulation (EU) 2026/1744, does not regulate "NLP" as a single field, nor does it classify every large language model (LLM) application as high-risk. It follows a risk-based and role-based approach. Its different gateways are use-specific: intended purpose, deployment context, affected people, and influence over consequential decisions are central to many Annex III uses and to Article 6(3), but prohibited practices and Article

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

50 transparency duties have their own tests (European Parliament and Council of the European Union 2024, 2026).

That distinction is easy to lose in technical discussion. We often speak as though a model and a deployed system were interchangeable. Under the Act, however, a general-purpose AI model is an upstream component that can be integrated into many systems. The resulting systems may fall into very different legal categories. The Act separately defines providers of AI systems and providers of general-purpose AI models; a deployer uses an AI system under its authority outside a personal non-professional activity; and a downstream provider may integrate a general-purpose model into a specific system. The same organisation can occupy more than one role. Distributed responsibility therefore means role-specific duties along a value chain, not automatically joint or undifferentiated responsibility (European Parliament and Council of the European Union 2024, Art. 3).

For Responsible NLP, this produces the following methodological proposition: *context is part of the system*. A benchmark cannot tell us whether a summariser will be used to shorten a personal document, prepare a medical report, assist a judicial authority in researching and interpreting facts and law, or influence an asylum decision. However, those contexts determine the severity of error, the power imbalance between actors, the possibility of contesting an output, and the safeguards that are needed.

The central methodological argument of this Responsible NLP piece is that responsibility must be demonstrated through an evidence chain connecting the model, the task-specific system, the institutional workflow, and the people affected by its use.

2. Three layers of responsibility

The Act becomes easier to understand if we separate three objects that are often collapsed in NLP discourse: the general-purpose model, the application built around it, and the institution that uses the application. Each layer produces different evidence and different duties.

At the ***model layer***, a provider develops or places a general-purpose model on the market. The model may support summarisation, translation, classification, dialogue, and code generation without being designed for one consequential decision. Chapter V therefore focuses on upstream documentation, information for downstream providers, copyright policy, and a public summary of training content. Where a model presents systemic risk, the duties expand to evaluation, adversarial testing, systemic-risk mitigation, incident reporting, and cybersecurity. Article 51 allows designation on the basis of high-impact capabilities or a Commission decision; the greater-than-10^{25}-FLOP training-compute threshold creates a presumption, not the only route to designation. This layer asks what capabilities and limitations travel with a reusable model.

At the ***system layer***, a downstream provider combines that model with prompts, retrieval, task-specific data, rules, interfaces, and decision logic. Intended purpose now becomes decisive. A general-purpose model integrated into a hiring system may become part of a high-risk AI system, while the same model used to rewrite internal emails may not. The relevant evidence must consequently concern the assembled system and its actual task, not only the base model.

At the ***institutional layer***, a deployer introduces the system into an organisation with its own authority structures, incentives, time pressures, professional norms, and affected populations. A technically identical system can operate differently in two institutions. One municipality may use a chatbot only to retrieve published information; another may allow staff to treat its recommendation as a presumptive eligibility decision. Risk therefore arises not only from model error but from workflow design, organisational dependence, and unequal power between the institution and the person subject to its decision.

Table 1 summarises the three analytical layers and the corresponding Responsible NLP question at each level. The three layers, the two axes below, "evidence chain," and "decision-mediated impact" are the author's analytical constructs, not statutory classifications or terms in the Act.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

Table 1. Three layers for analysing responsibility in NLP systems

Layer	Object of analysis	Principal actors	Responsible NLP question
Model	General capabilities, training, limitations, and systemic risks	GPAI-model provider	What information and risk evidence must travel downstream?
System	Task-specific pipeline, data, interface, and intended purpose	AI-system provider and downstream provider	Is the assembled system valid and safe for this particular use?
Institution	Workflow, authority, affected people, oversight, and procedural accountability	Deployer and domain professionals	Can people recognise, prevent, and contest harm in practice?

This reveals two regulatory axes. The first runs *along the value chain*: model provider, downstream provider, deployer, and affected person. The second runs *through the lifecycle*: problem formulation, data construction, development, evaluation, deployment, monitoring, and retirement. Responsibility gaps appear when information fails to move across the first axis or harms are examined too late on the second. Article 25 addresses specified value-chain transitions: in defined circumstances, a distributor, importer, deployer, or other third party is treated as the provider of a particular high-risk system. As amended, it also requires relevant cooperation, information about known limitations and failure modes, reasonably expected technical access, and, where applicable, a written agreement between the high-risk-system provider and a third-party supplier (European Parliament and Council of the European Union 2024, 2026, Art. 25). Yet contracts and documentation cannot remove the deeper coordination problem: no single actor necessarily sees the complete sociotechnical system.

The three layers are therefore analytically distinct but operationally interdependent: evidence about capabilities and limitations must move downstream, while incidents, complaints, and experience from deployment must move upstream.

3. A risk map for language technologies

Before applying the risk map, a threshold question must be answered: does the NLP tool satisfy the definition of an AI system in Article 3(1), is it a general-purpose AI model, or does the deployment involve both an upstream model and a downstream AI system? This distinction matters for deterministic or exclusively rules-based language software and for systems assembled around a GPAI model. The next questions concern scope, actor roles, prohibitions, high-risk classification, transparency, model-level obligations, applicable dates, and other governing law (European Parliament and Council of the European Union 2024, Arts. 2–3).

An AI Act check for an NLP use.

1. Is the object an AI system, a GPAI model, or both a model and a downstream system?
2. Is it within the territorial and material scope, and does an exclusion apply?
3. What role does each actor occupy: model provider, AI system or downstream provider, deployer, importer, or distributor?
4. Does an Article 5 prohibition apply?
5. Is the system high-risk under Article 6 and Annex I or III, including the Article 6(3) derogation?
6. Does Article 50 apply at the interaction or output level?

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

7. Do Articles 53 or 55 apply at the model level?
8. Which duties are already applicable, and what other Union or national law governs?

The Act can be read as a map with several routes for NLP systems.

Prohibited practices. Article 5 prohibits a limited set of AI practices. For NLP, persuasion or deceptive language is not prohibited merely because it may influence a user. Article 5(1)(a) applies only where an AI system deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques; the objective or effect is materially to distort behaviour by appreciably impairing informed decision-making; the person is thereby caused to take a decision that they would not otherwise have taken; and significant harm is caused or reasonably likely. Article 5(1)(b) separately addresses exploitation of vulnerability due to age, disability, or a specific social or economic situation, subject to its own material-distortion, decision, and significant-harm conditions. Emotion recognition based on biometric data in workplaces and educational institutions is generally prohibited, subject to the medical- or safety-purpose exception. A text-only sentiment classifier is therefore not automatically the same legal category. Regulation (EU) 2026/1744 also added prohibitions concerning specified non-consensual intimate or sexually explicit material and child-sexual-abuse material; those provisions apply from 2 December 2026. The Commission's prohibited-practices guidelines are non-binding interpretive guidance (European Parliament and Council of the European Union 2024, 2026; European Commission 2025a, Arts. 3(39) and 5).

High-risk systems. NLP-relevant Annex III fields include education, employment, access to essential private services and essential public assistance benefits and services, law enforcement, migration and asylum, and the administration of justice. Many listed activities depend heavily on text classification, information extraction, ranking, translation, summarisation, or question answering. A system intended to filter job applications, assess learning outcomes, evaluate eligibility for public assistance benefits, or assist a judicial authority in researching and interpreting facts and law and applying the law to a concrete set of facts can therefore be high-risk even when its NLP components look ordinary. Purely ancillary administrative activities are excluded from the justice listing (European Parliament and Council of the European Union 2024, Art. 6 and Annex III).

The classification is not mechanical. Article 6(3) can exclude an Annex III use only where the system does not pose a significant risk of harm to health, safety, or fundamental rights, including by not materially influencing the outcome of decision-making, and it performs one of four functions: a narrow procedural task; improvement of the result of a previously completed human activity; detection of decision-making patterns or deviations without replacing or influencing the completed human assessment absent proper human review; or a preparatory task for an assessment relevant to an Annex III use. Providers relying on this provision must document the assessment and register the system under Article 49(2). An Annex III system that profiles natural persons remains high-risk. The question is therefore not merely whether an LLM appears somewhere in a workflow, but what influence it has over the outcome (European Parliament and Council of the European Union 2024, Arts. 6(3)–(4) and 49(2)).

This exception creates an important boundary problem for NLP. A provider may describe a system as a decision-support tool or a preparatory summariser, but a summary can determine which evidence a human notices, and a ranking can define the practical set of candidates considered. Formal decision authority and actual influence are not the same. Researchers should therefore measure *decision-mediated impact*: how often humans follow, reject, or fail to notice model suggestions; whether omissions propagate into final decisions; and whether the interface makes alternatives realistically available. Classification should be informed by empirical workflow evidence rather than product labels alone.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

Article 50 transparency duties (an explanatory grouping). Article 50 transparency duties (an explanatory grouping). Article 50 does not create a statutory "transparency-risk" tier. It allocates particular duties between actors. Providers must design systems that interact directly with people so users are informed that they are interacting with AI unless this is obvious to a reasonably well-informed, observant, and circumspect person, subject to the stated law-enforcement exception. Providers of systems generating synthetic audio, image, video, or text must enable effective machine-readable marking and detection as artificially generated or manipulated, as far as technically feasible. Article 50(2) states separate exceptions where a system performs an assistive function for standard editing or does not substantially alter the input data or its semantics. Deployers publishing AI-generated or manipulated text to inform the public on matters of public interest must disclose its artificial origin unless human review or editorial control has occurred and a natural or legal person holds editorial responsibility. Article 50 applies from 2 August 2026; the limited transition until 2 December 2026 concerns Article 50(2) systems placed on the market before 2 August 2026, not the other Article 50 duties (European Parliament and Council of the European Union 2024, 2026, Art. 50).

General-purpose AI models. Chapter V creates obligations for providers of GPAI models. Article 53 requires technical documentation, information for downstream providers, an EU-copyright-compliance policy, and a public summary of training content following the AI Office template; it does not require publication of the complete training corpus. Under Article 53(2), a qualifying free- and open-source provider is exempt only from Article 53(1)(a)–(b): the copyright policy and public summary remain, and the exemption does not cover a GPAI model with systemic risk. Providers of models with systemic risk face the additional Article 55 duties, including evaluation, adversarial testing, systemic-risk assessment and mitigation, serious-incident reporting, and cybersecurity. Chapter V began applying on 2 August 2025, while covered GPAI models placed on the market before that date have until 2 August 2027. The voluntary GPAI Code of Practice offers a non-binding route for demonstrating compliance; it does not amend the Regulation or conclusively establish compliance (European Parliament and Council of the European Union 2024; European Commission 2025b, Arts. 51, 53, 55, 111 and 113).

Uses outside the prohibited- and high-risk regimes (an explanatory grouping). Uses outside the prohibited- and high-risk regimes (an explanatory grouping). A grammar checker, meeting-note summariser, search aid, or low-stakes translation tool may not fall into a prohibited or high-risk category. That does not mean it is harmless, and it does not remove potential Article 4 or Article 50 duties, upstream GPAI obligations, or obligations under data protection, copyright, consumer protection, employment, or other law. It means that the AI Act itself imposes fewer system-specific duties. Responsible NLP must remain broader than legal classification.

4. Four NLP cases

The following four cases illustrate how different AI Act gateways apply to common NLP applications. Recruitment and educational-assessment systems can fall within Annex III because their outputs may significantly affect access to employment or education. A public-information chatbot may engage Article 50 without being high-risk, whereas a system that evaluates eligibility for essential public assistance benefits or services can fall within Annex III point 5(a). AI-generated public-interest text is addressed through Article 50's disclosure and provenance duties. For the examples below, the high-risk requirements in Chapter III, Sections 1–3 will apply to Article 6(2)/Annex III systems from 2 December 2027; the present-tense descriptions identify the enacted duties rather than implying that all are already applicable on 1 August 2026 (European Parliament and Council of the European Union 2024, 2026).

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

4.1 A recruitment assistant

Consider an LLM-based system that extracts skills from CVs, ranks applicants, and drafts recommendations for recruiters. A system intended for recruitment or candidate selection, particularly to analyse and filter applications or evaluate candidates, falls within Annex III point 4(a) and, subject to Article 6(3), will be high-risk once the relevant provisions apply. Its provider will need a lifecycle risk-management process, compliant training, validation, and testing data where applicable, technical documentation, logging, deployer information, human-oversight measures, and suitable accuracy, robustness, and cybersecurity. Under Article 26(7), an employer deploying a high-risk system in the workplace must first inform worker representatives and affected workers that they will be subject to its use (European Parliament and Council of the European Union 2024, Arts. 6, 9–15, 26(7), and Annex III(4)).

This case shows why average accuracy is an inadequate measure. CV language varies with profession, age, disability, educational opportunity, migration history, dialect, and native-language background. A model may treat the vocabulary of one social group as a proxy for competence while discounting another group's experience. For a high-risk system that uses model-training techniques, Article 10 requires the relevant training, validation, and testing datasets to address data origin, collection processes, annotation, assumptions, representativeness, geographical and contextual conditions, data gaps, and biases likely to affect fundamental rights. This aligns with long-standing NLP arguments for data statements and explicit reasoning about who is represented, who is harmed, and why (Bender and Friedman 2018; Blodgett *et al.* 2020).

Even a demographically balanced dataset may preserve discrimination if its labels encode an unequal past. If successful employment is inferred from historical performance ratings, promotion, or retention, the target variable may reflect inaccessible workplaces, managerial bias, or unequal caring responsibilities rather than job competence. Adding more records from underrepresented groups would then reproduce the same measurement error at greater scale. The distinction between representation bias and label or measurement bias is essential: Article 10 creates a governance obligation, but technical teams still need a defensible theory of what their labels mean (Suresh and Guttag 2021).

Human review is not a magic remedy. Article 14 primarily requires providers to design and develop high-risk systems so that natural persons can exercise effective oversight, including awareness of automation bias and, as appropriate, the ability to interpret, disregard, override, reverse, or interrupt outputs or operation. Article 26 requires deployers to assign oversight to people with the necessary competence, training, authority, and support. A recruiter who clicks "approve" beside a model ranking without time, authority, or evidence is not meaningful human oversight. Responsible NLP research can contribute calibrated uncertainty, counterfactual explanations, subgroup testing, interpretable evidence displays, and interfaces designed to support genuine disagreement with the model (European Parliament and Council of the European Union 2024, Arts. 14 and 26).

Meaningful oversight therefore has at least four conditions: the reviewer needs relevant competence, sufficient information, real authority to intervene, and enough time to exercise judgment. It also needs organisational protection. If disagreement with the model lowers productivity scores or requires lengthy justification, the interface formally permits an override while the institution effectively discourages it. Evaluating oversight means studying behaviour and incentives, not merely confirming the presence of a human in the process.

4.2 An educational system

Now consider three systems used by a university. A chatbot answers questions about library opening hours. An automated essay scorer influences grades. An online-proctoring system attempts to detect prohibited behaviour during an examination.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

For the first system, Article 50(1) requires the provider to design it so that students are informed they are interacting with AI unless this is obvious under the statutory standard. A scorer intended to evaluate learning outcomes and a system intended to monitor prohibited student behaviour during a test fall within Annex III points 3(b) and 3(d), subject to Article 6(3). If a system infers emotions from biometric data in an educational institution, its use is generally prohibited rather than merely high-risk, subject to the medical- or safety-purpose exception (European Parliament and Council of the European Union 2024, Arts. 5(1)(f), 50(1), and Annex III(3)).

The technical lesson is that "education NLP" is not one risk category. We need evaluations tied to use and consequence. A tutoring chatbot should be tested for hallucination, pedagogical appropriateness, accessibility, age-related vulnerability, and multilingual behaviour. An essay scorer additionally requires evidence of validity across subjects, languages, writing styles, and disability-related forms of expression. A proctoring system requires analysis of false accusations, surveillance effects, and routes for appeal. The model may be impressive in a laboratory and still be unacceptable in a classroom.

The essay-scoring example also exposes a construct-validity problem. Grammar, lexical variety, and conformity to a dominant rhetorical style are measurable, but they are not identical to reasoning, learning, or creativity. Optimising a model against convenient textual proxies can narrow what students are rewarded for writing. Once students adapt to the scorer, the system also changes the behaviour it was designed to measure. Responsible evaluation should therefore include educational validity, feedback effects, and the possibility that some learning goals should remain outside automated assessment.

4.3 A public-service chatbot

A municipal chatbot that only explains how to apply for housing assistance is not automatically high-risk; where Article 50(1) applies, its provider must enable the required interaction notice. If the same interface is intended by a public authority or on its behalf to evaluate a person's eligibility for essential public assistance benefits and services, including healthcare services, it can fall within Annex III point 5(a). Use by a public authority, or influence over a generic public decision, is not by itself enough to trigger that particular category (European Parliament and Council of the European Union 2024, Art. 50(1) and Annex III(5)(a)).

Once the relevant provisions apply on 2 December 2027, obligations will extend beyond the model provider. Public bodies and private entities providing public services must conduct a fundamental-rights impact assessment (FRIA) before deploying covered Annex III systems. Deployers must monitor operation, keep logs under their control for at least six months unless other law provides otherwise, and provide required information to affected people. Under Article 26(5), a deployer must suspend use where it has reason to consider that use in accordance with the instructions may present a risk within Article 79(1), while a separately defined serious incident triggers an additional reporting path (European Parliament and Council of the European Union 2024, 2026, Arts. 26–27, 79(1), and 113).

Affected people also gain procedural tools. Article 85 permits a natural or legal person with grounds to consider that the Act has been infringed to complain to the relevant market-surveillance authority. Article 86 creates a qualified right to explanation where a person is subject to a decision taken by a deployer on the basis of output from an Annex III high-risk system, other than a point 2 critical-infrastructure system, and the decision produces legal effects or similarly significantly affects that person in a way they consider adverse to health, safety, or fundamental rights. Subject to statutory exceptions and to rights already provided by Union law, the person may obtain from the deployer a clear and meaningful explanation of the system's role in the procedure and the main elements of the decision. This is not a universal right to inspect model internals; it is a context-bound mechanism for understanding and challenging

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

consequential decision-making (European Parliament and Council of the European Union 2024, Arts. 85–86). The distinction should shape XAI research: an explanation is valuable not because it visualises attention weights, but because it helps a person understand, question, and challenge an outcome.

The FRIA is especially important because some risks cannot be determined by the provider in advance. The deployer knows which population will encounter the system, what alternatives exist, how errors are corrected, and whether refusing automated assistance carries a penalty. Under amended Article 27(4), when a data-protection impact assessment already meets part of the FRIA obligations, the deployer may cross-reference the relevant sections or incorporate relevant parts. The instruments remain overlapping but non-identical: privacy is crucial, while fundamental-rights analysis can additionally cover discrimination, access to services, expression, dignity, and effective remedy (European Parliament and Council of the European Union 2024, 2026, Art. 27(4)).

4.4 AI-generated public-interest text

Article 50 is particularly important for journalism, public communication, and fact-checking. Under Article 50(2), a provider of a generative system must enable synthetic outputs to be marked in a machine-readable format and detectable as artificially generated or manipulated, as far as technically feasible. Under Article 50(4), a deployer publishing AI-generated or manipulated text to inform the public on matters of public interest must disclose that origin, unless human review or editorial control has occurred and a natural or legal person holds editorial responsibility. Provider-side machine-readable marking does not replace the deployer's human-perceivable disclosure where Article 50(4) applies (European Parliament and Council of the European Union 2024, Art. 50(2) and (4)).

This creates two different forms of transparency. The provider supports technical provenance at the output layer; the publisher provides a perceptible disclosure to the reader. The Commission's July 2026 guidelines are non-binding interpretive guidance. The Code of Practice on Transparency of AI-Generated Content is a voluntary implementation route. Neither source amends the Regulation or conclusively establishes compliance (European Commission 2026a, 2026b).

But a label saying "AI-generated" does not establish that a text is false, and the absence of a label does not establish that it is true. Provenance is not veracity. Watermarks and metadata can be removed, degraded by translation or paraphrasing, or lost across platforms. Responsible NLP therefore needs robust provenance methods, but also source attribution, factuality evaluation, uncertainty communication, editorial workflows, and public literacy.

5. From principles to evidence

The Act's most important contribution to Responsible NLP may be its conversion of broad principles into work products. Once the relevant high-risk provisions apply, Article 9 requires lifecycle risk management; Article 10 governs relevant training, validation, and testing datasets; Articles 11 and 12 require technical documentation and logging; Articles 13 and 14 connect information for deployers with human-oversight design; and Article 15 requires accuracy, robustness, and cybersecurity. These are principally requirements for high-risk-system providers, complemented by deployer duties in Article 26 (European Parliament and Council of the European Union 2024, Arts. 9–15 and 26).

For Responsible NLP, the important shift is evidentiary: responsibility is not a property that a model possesses, but a claim that must be supported by connected evidence across the model, system, and institutional layers.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

Table 2. An illustrative crosswalk from legal hooks to NLP and institutional evidence

Legal hook	Illustrative evidence	Primary owner
Article 9	Risk register, foreseeable-misuse scenarios, and residual-risk decisions	AI-system provider
Article 10	Data statement, label-validity analysis, and subgroup and multilingual evaluation	Provider and data team
Articles 11–13	System card, logs, deployer instructions, and known limitations	Provider
Articles 14 and 26	Design and workflow studies showing competence, time, support, authority, and actual override behaviour	Provider and deployer
Article 15	Accuracy, robustness, security, and distribution-shift evidence	Provider
Article 27	Population-, workflow-, complaint-, and rights-specific FRIA	Covered deployer
Article 50	Interaction notice, provenance marking, and public-disclosure workflow	Provider or deployer, by paragraph
Article 72	Monitoring plan, incident and complaint triggers, and corrective-action rules	Provider, with deployer feedback

These obligations resemble familiar research artefacts: data statements, model cards, evaluation reports, audit trails, red-team findings, uncertainty estimates, interface studies, and post-deployment monitoring. The difference is that documentation can no longer be treated as an optional appendix added after the interesting modelling work. It becomes part of whether a system is suitable for use.

The Act also challenges the benchmark culture of NLP. It repeatedly ties requirements to the *intended purpose*, foreseeable misuse, affected groups, and deployment context. This is compatible with research showing that language-model harms span discrimination, misinformation, malicious use, human-computer interaction, privacy, and environmental and socioeconomic effects (Weidinger *et al.* 2022). It also supports end-to-end auditing: harms can emerge during problem formulation, data construction, model development, integration, interface design, procurement, or institutional use (Raji *et al.* 2020).

The deeper shift is from claiming that a system is responsible to producing an *evidence chain*. For a high-risk system, evidence should connect the intended purpose to foreseeable harms; harms to evaluation criteria; criteria to data and test design; results to residual-risk decisions; and those decisions to instructions, oversight, monitoring, and corrective action. A fairness table that is disconnected from the deployment population is weak evidence. A model card that documents the base model but not the retrieval corpus, prompt layer, or institutional workflow is incomplete evidence. Table 2 translates this evidence-chain perspective into an illustrative crosswalk between AI Act provisions and the corresponding NLP and institutional evidence.

For high-risk-system providers, this chain enters the quality-management system, conformity assessment, registration, and post-market monitoring. Conformity assessment is mainly a pre-market mechanism and must be repeated following a substantial modification, while Article 72 requires providers to collect and analyse performance data throughout the system's lifetime (European Parliament and Council of the European Union 2024, Arts. 17, 43, 49, and 72). This matters for NLP because language, user behaviour, institutions, and information environments change. A model may remain numerically unchanged while its error distribution or social meaning changes. Monitoring therefore needs trigger conditions: which shifts, incidents, complaints, or subgroup disparities require investigation, restriction, retraining, or withdrawal?

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

Red teaming must also be connected to who can be harmed. Generic jailbreak testing is useful but insufficient for a multilingual public-service assistant or a recruitment tool. Sociotechnical red teaming can recruit participants with relevant linguistic and social knowledge, parameterise tests around specific harms, preserve disagreement rather than treating it as annotation noise, and use arbitration to understand contested judgments (Weidinger *et al.* 2024). The research contribution is not only finding more unsafe prompts; it is developing defensible coverage of a system's risk surface.

6. The AI Act does not stand alone

The AI Act is cumulative rather than self-contained. Article 2(7) preserves Union rules on personal-data protection, privacy, and communications confidentiality (European Parliament and Council of the European Union 2024, 2026, Art. 2(7)). An NLP system can therefore satisfy its AI Act classification duties and still process personal data unlawfully under the General Data Protection Regulation. Conversely, a lawful basis for processing does not establish that an automated decision system is accurate, non-discriminatory, or appropriate for its institutional purpose.

This interaction is concrete for language models. Training corpora, retrieval stores, prompts, conversation logs, and generated outputs may contain personal data. The European Data Protection Board has stressed that anonymity must be assessed case by case and that, with reasonable means, both extraction of training-related personal data and obtaining information about the relevant data subjects through queries must be insignificant (European Data Protection Board 2024, paras. 34, 38 and 43). Whether data were publicly accessible is only one factor in assessing reasonable expectations, not an automatic permission to process them (European Data Protection Board 2024, paras. 91–95). Unlawful development-stage processing may affect the lawfulness of later processing, depending on a case-specific analysis (European Data Protection Board 2024, paras. 121–132).

Bias evaluation creates a further tension. Detecting unequal performance may require sensitive demographic data, while processing special categories of personal data is restricted. Regulation (EU) 2026/1744 deleted former Article 10(5) and inserted Article 4a. Article 4a(1) permits high-risk-system providers to process such data exceptionally for bias detection and correction only where strictly necessary and where detailed safeguards are satisfied, including inability effectively to use other data, technical limits on reuse, privacy-preserving and security measures, access controls, non-disclosure, deletion, and documentation. Article 4a(2) can also apply, under additional harm- and discrimination-related conditions, to providers and deployers of other AI systems and models and to deployers of high-risk systems. It creates neither a duty to collect demographic attributes nor a general exemption from data-protection law. Responsible NLP therefore needs privacy-preserving evaluation, necessity and proportionality analysis, community governance, and methods for identifying disparities where direct labels are unavailable or inappropriate (European Parliament and Council of the European Union 2024, 2026, Art. 4a(1)–(2)).

Other legal regimes may also govern the same system: employment and equality law for recruitment, consumer law for conversational services, copyright law for training and generation, sectoral rules for health or finance, and the Digital Services Act for certain platform practices. The important methodological lesson is that risk categories are not moral labels. A system that is not classified as high-risk under the AI Act is not necessarily socially low-risk, and a CE marking concerns conformity of the relevant high-risk system with applicable requirements; it does not certify that every output is truthful or fair. Compliance answers defined legal questions; Responsible NLP must also investigate harms that the classification system does not capture.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

7. What regulation does not solve

The AI Act is significant, but compliance should not be confused with complete responsibility.

The fundamental difficulty is sociotechnical. NLP research often draws the system boundary around data, a model, and an output metric, while fairness and accountability depend on institutions, histories, incentives, and the meaning of the task itself. This is the abstraction trap: a technically coherent intervention can fail when its simplified representation of a social problem is mistaken for the problem (Selbst *et al.* 2019). The Act pushes researchers toward context, but it cannot supply the missing contextual knowledge.

First, the Act does not establish multilingual or cultural parity. Its data-governance rules require relevant and sufficiently representative data and attention to geographical and contextual settings for high-risk systems. These are valuable hooks for multilingual evaluation. They do not specify how to measure semantic fidelity across languages, protect dialects, govern Indigenous or community data, or decide when a system should not be deployed because local evidence is inadequate.

Second, transparency does not guarantee understanding. Informing someone that they are speaking with a chatbot is useful, but it does not communicate hallucination rates, uncertainty, knowledge limits, or possible disparities. Likewise, machine-readable marking of generated text does not guarantee durable detection. The 2025 Foundation Model Transparency Index found persistent opacity concerning training data and post-deployment impact, illustrating the distance between formal disclosure and meaningful public knowledge (Wan *et al.* 2025).

Third, the right to explanation is deliberately limited. Many NLP systems that influence people will fall outside Article 86. Even within its scope, useful explanations depend on the affected person's goal: correcting data, understanding a rule, identifying discrimination, or appealing a decision. XAI should therefore be evaluated by actionability and contestability, not only fidelity to model mechanics.

Fourth, the research exclusions need careful reading. Article 2(6) excludes AI systems or models, including their outputs, specifically developed and put into service for the sole purpose of scientific research and development. Article 2(8) generally excludes research, testing, or development activity before placement on the market or putting into service, but preserves applicable Union law and does not cover real-world testing. Recital 25 adds that research and development should follow recognised ethical and professional standards for scientific research. Neither provision creates a blanket exemption for a university system deployed operationally. More importantly, legal exclusion is not ethical exemption. Research can still create privacy, labour, representational, environmental, and downstream harms (European Parliament and Council of the European Union 2024, Art. 2(6), Art. 2(8), and recital 25).

Finally, the Act is a floor. An NLP system may fall outside Article 5 and high-risk classification while still marginalising a low-resource language, reinforcing stereotypes, exposing annotators to traumatic content, or encouraging over-reliance. Responsible NLP should ask a wider question than "Are we compliant?": *Is this system worthy of the trust and power it is being given?*

8. A research agenda for the NLP community

The Act creates concrete opportunities for NLP research. The layer labels below are not rigid categories; they indicate where the primary burden of evidence falls while preserving the cross-layer nature of most Responsible NLP problems:

- **System layer—Use-sensitive evaluation** replace one global score with an evaluation argument connecting intended purpose, affected groups, foreseeable misuse, severity of failure, and institutional conditions. Research should test whether benchmark conclusions survive realistic workflows and distribution shifts.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

- **Model and system layers—Multilingual and cultural assurance:** develop methods for identifying data gaps, measuring cross-language disparities, and distinguishing translation quality from culturally appropriate communication. A valid outcome may sometimes be a documented decision not to deploy in a language for which evidence is inadequate.
- **System and institutional layers—XAI, uncertainty, and contestability:** compare explanations by what users can do with them: detect an error, correct data, understand a governing rule, challenge discrimination, or appeal a decision. Confidence displays should be evaluated for calibration and for their effect on automation bias.
- **Model and system layers: Provenance and content integrity:** improve robust marking, metadata preservation, source attribution, and detection while studying degradation through paraphrasing, translation, screenshots, and platform transformations. Provenance claims should state what they establish and what they do not.
- **Institutional layer—Human-oversight systems:** test whether people can recognise failure, obtain contrary evidence, override outputs, and exercise authority under realistic time and organisational constraints. The unit of evaluation should be the human–AI workflow rather than the model alone.
- **Institutional layer and lifecycle—Post-deployment evaluation:** create monitoring methods, incident taxonomies, logging protocols, and feedback processes that reveal harms not visible before release. This includes methods for detecting feedback loops and performance changes without subjecting affected people to excessive surveillance.
- **Cross-layer—Compliance-aware documentation:** connect data statements, model cards, system cards, impact assessments, conformity evidence, and audit records across the model-to-system value chain. Research can examine whether downstream actors actually receive enough information to fulfil their duties.
- **Institutional and cross-layer—Participatory risk discovery:** involve affected communities, domain professionals, annotators, and civil-society organisations before evaluation categories are fixed. Participation should influence problem formulation and deployment decisions, not merely provide labels for a predefined benchmark.

These are not peripheral "ethics topics." They are methods, datasets, metrics, interfaces, protocols, and empirical studies. The Act makes their practical relevance unusually clear.

One further research object is *regulatory translation*: turning terms such as appropriate, sufficiently representative, meaningful, and reasonably foreseeable into testable but context-sensitive evidence. A universal metric will rarely settle these concepts. Better work will combine quantitative evaluation, qualitative inquiry, legal interpretation, and domain expertise while making disagreements and assumptions visible.

9. A moving implementation landscape

The legal position in this piece is current to 1 August 2026. Regulation (EU) 2026/1744, the Digital Omnibus on AI, entered into force on 27 July 2026 and amended both the implementation timetable and several substantive provisions. Chapters I and II began applying on 2 February 2025. Article 4 has applied since that date, but its amended wording took effect on 27 July 2026; the original Article 5 prohibitions have applied since 2 February 2025. Governance and GPAI-model provisions began applying on 2 August 2025, subject to the separate 2 August 2027 transition for covered GPAI models placed on the market before 2 August 2025 (European Parliament and Council of the European Union 2024, 2026).

Under Article 113's general rule, the remaining provisions apply from 2 August 2026 unless a specific transition or later date governs. Article 50 applies from that date, although providers of

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

systems covered by Article 50(2) and placed on the market before it have until 2 December 2026 to comply with that paragraph. The new Article 5(1)(ba)–(bb), (1a), and (1b) provisions also apply from 2 December 2026. Chapter III, Sections 1–3 apply to Article 6(2)/Annex III high-risk systems from 2 December 2027 and to Article 6(1)/Annex I high-risk systems from 2 August 2028. Article 111 contains separate legacy-system rules, including a 2 August 2030 compliance date for covered high-risk systems intended for use by public authorities. These dates distinguish entry into force from entry into application: an obligation may be enacted law before it applies to a particular actor or system (European Parliament and Council of the European Union 2024, 2026, Arts. 111 and 113).

AI literacy deserves more attention than it usually receives. As amended, Article 4 requires providers and deployers to take measures supporting the development of AI literacy among staff and other persons dealing with the operation and use of AI systems on their behalf. Measures must take account of technical knowledge, experience, education and training, the context of use, and the persons or groups on whom the systems are used. The provision does not require a provider or deployer to guarantee any specific level of literacy for an individual. For Responsible NLP, appropriate measures should nevertheless be role-specific and operational: a recruiter needs to understand ranking limitations and automation bias; a teacher needs to understand validity and student adaptation; and a public official needs to understand when an output affects rights and how it can be challenged. The Commission's practical examples can assist, but copying an example does not itself establish compliance (European Parliament and Council of the European Union 2024, 2026; European Commission 2025c, Art. 4).

Readers should also distinguish legal sources. Regulations (EU) 2024/1689 and 2026/1744 are binding law, subject to staggered application dates. Commission guidelines are non-binding interpretation. Codes of Practice are voluntary compliance routes and evidence, not legislation or conclusive proof of compliance. Harmonised standards, once referenced in the Official Journal, may provide a presumption of conformity only to the extent that they cover an obligation. Responsible NLP frameworks can supply valuable scientific and normative evidence, but do not themselves establish legal compliance.

The dates, consolidated statutory text, and implementation guidance in this section should therefore be checked again at proof stage.

This moving landscape is itself a lesson for interdisciplinary NLP. Responsible deployment requires collaboration among computer scientists, lawyers, social scientists, domain professionals, affected communities, and public authorities. No discipline can interpret intended purpose, fundamental-rights risk, technical feasibility, and real-world harm alone.

10. Conclusion

The EU AI Act does not provide an algorithm for Responsible NLP. What it provides is a new baseline: language technologies must be understood through their purpose, context, lifecycle, value chain, and effects on people.

The same model can be a low-stakes writing aid, the engine of a transparent chatbot, a component of a high-risk recruitment system, or part of a consequential public decision. The responsible question is therefore no longer simply, "How capable is this model?" It is, "What role is this system being given, what can happen when it fails, and who has the knowledge, authority, and means to respond?"

Law can require documentation, oversight, transparency, and complaint, explanation, and procedural-accountability mechanisms, supplemented by remedies under other Union and national law. It cannot by itself make an explanation meaningful, a dataset culturally representative, a human reviewer independent, or a multilingual system fair. That is where the research agenda begins. The AI Act should not be the endpoint of Responsible NLP. It should be a reason to make responsibility more empirical, more interdisciplinary, and more central to the field.

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.

Competing interests. The author declare that the manuscript complies with the ethical standards of the journal and that there are no competing interests to disclose.

References

Bender E.M. and **Friedman B.** (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* **6**, 587–604. https://doi.org/10.1162/tacl_a_00041.

Blodgett S.L., **Barocas S.**, **Daumé H.**III and **Wallach H.** (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

European Commission (2025a). *Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689*, February. Available at https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act (accessed 1 August 2026).

European Commission (2025b). *Repository of AI Literacy Practices*, November. Available at https://digital-strategy.ec.europa.eu/en/policies/repository-ai-literacy-practices (accessed 1 August 2026).

European Commission (2025c). *The General-Purpose AI Code of Practice*, July. Available at https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (accessed 1 August 2026).

European Commission (2026a). *Code of Practice on Transparency of AI-Generated Content*, June. Available at https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content (accessed 1 August 2026).

European Commission (2026b). *Guidelines on the Implementation of the Transparency Obligations for Certain AI Systems under Article 50 of the AI Act*, July. Available at https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems (accessed 1 August 2026).

European Data Protection Board (2024). *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models*, December. Available at https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en (accessed 1 August 2026).

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*. L 2024/1689, 12 July 2024. http://data.europa.eu/eli/reg/2024/1689/oj.

European Parliament and Council of the European Union (2026). Regulation (EU) 2026/1744 of 8 july 2026 amending regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital omnibus on AI). *Official Journal of the European Union*. L 2026/1744, 24 July 2026, July 8, 2026. http://data.europa.eu/eli/reg/2026/1744/oj (accessed 1 August 2026).

Raji I.D., **Smart A.**, **White R.N.**, **Mitchell M.**, **Gebru T.**, **Hutchinson B.**, **Smith-Loud J.**, **Theron D.** and **Barnes P.** (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, pp. 33–44. https://doi.org/10.1145/3351095.3372873.

Selbst A.D., **Boyd D.**, **Friedler S.A.**, **Venkatasubramanian S.** and **Vertesi J.** (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, pp. 59–68. https://doi.org/10.1145/3287560.3287598.

Suresh H. and **Guttag J.** (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Association for Computing Machinery, pp. 1–9. https://doi.org/10.1145/3465416.3483305.

Wan A., **Klyman K.**, **Kapoor S.**, **Maslej N.**, **Longpre S.**, **Xiong B.**, **Liang P.** and **Bommasani R.** (2025). The 2025 foundation model transparency index. Stanford Center for Research on Foundation Models. Available at https://crfm.stanford.edu/fmti/December-2025/paper.pdf (accessed 1 August 2026).

Weidinger L., **Mellor J.F.J.**, **Pegueroles B.G.**, **Marchal N.**, **Kumar R.**, **Lum K.**, **Akbulut C.**, **Diaz M.**, **Bergman A.S.**, **Rodriguez M.D.**, **Rieser V.** and **Isaac W.** (2024). STAR: Sociotechnical approach to red teaming language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 21516–21532. https://doi.org/10.18653/v1/2024.emnlp-main.1200.

Weidinger L., **Uesato J.**, **Rauh M.**, **Griffin C.**, **Huang P.-S.**, **Mellor J.**, **Glaese A.**, **Cheng M.**, **Balle B.**, **Kasirzadeh A.**, **Biles C.**, **Brown S.**, **Kenton Z.**, **Hawkins W.**, **Stepleton T.**, **Birhane A.**, **Hendricks L.A.**, **Rimell L.**, **Isaac W.**, **Haas J.**, **Legassick S.**, **Irving G.** and **Gabriel I.** (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, pp. 214–229. https://doi.org/10.1145/3531146.3533088.

Cite this article: Lamsiyah S (2026). The same model, different responsibilities: Responsible NLP under the EU AI Act. *Natural Language Processing* **32**, 662–675. https://doi.org/10.1017/nlp.2026.10036

Downloaded from https://www.cambridge.org/core. 17 Nov 2026 at 11:34:39, subject to the Cambridge Core terms of use.