

Journalism as Data: GDPR Implications of Licensing Journalistic Content to Large Language Models

Author(s) Gionata Bouché, Marten Steketee

Contact g.bouche@uva.nl, m.c.steketee@uva.nl

Affiliation(s) Gionata Bouché is a PhD candidate at the University of Amsterdam's Institute for Information Law (IViR) and at the Responsible Media Lab (REM).
Marten Steketee is a research assistant for the AMdEx project, University of Amsterdam.

Keywords Generative AI, News, GDPR, lawfulness, transparency, Media

Published

Received: 20 Nov. 2025

Accepted: 16 Mar. 2026

Published: 22 Aug. 2026

Citation

Gionata Bouché, Marten Steketee, Journalism as Data: GDPR Implications of Licensing Journalistic Content to Large Language Models, Technology and Regulation, 2026, 101-121 • 10.71265/7yhpdx81 • ISSN: 2666-139X

Abstract

This article explores the data protection implications under the GDPR of integrating journalistic content into Large Language Models (LLMs). The number of commercial partnerships between AI companies and news publishers for the licensing of daily news and archival content has rapidly increased. We contend that while LLMs and their top-layer applications do offer innovative solutions for news dissemination, publishers should carefully evaluate their position under the GDPR. Comparing different technical solutions available, in particular pre-training, fine-tuning and Retrieval Augmented Generation (RAG), we analyse the relevant regulatory barriers and opportunities, focusing in particular on the distribution of processing roles, lawfulness and transparency of these deals, and the application of the special regime for journalistic processing under art 85(2) GDPR.

1. Introduction

1.1 The evolving relationship between news publishers and LLM-providers

In December 2023, the New York Times decided to bring legal action against OpenAI and partner company Microsoft for the infringement of the newspapers' copyright vested in its news archival content.¹ The central claim to the litigation is that OpenAI allegedly trained its Large Language Models (LLMs), on which ChatGPT and other AI systems run, on millions of the Times' articles without its consent.² According to the newspaper's publisher and owner, AG Sulzberger, publishers like the Times, should receive fair compensation for their journalistic investments, especially from a company seeking to free-ride on its assets to build products and services potentially competing with the news organisations themselves.³

Despite the buzz surrounding the Times' lawsuit, relatively few publishers have decided to follow suit and escalate it to the courtroom.⁴ Quite the contrary, news organisations, including many European ones, have concluded strategic deals with OpenAI and other AI companies to settle (potential) conflicts. The list, which includes several European publishers, has been noticeably expanding.⁵

These developments have extensively engaged scholars and practitioners in intellectual property law. However, comparatively little attention has been given to the questions that the mass-scale processing of personal data contained in journalistic content such as latest news and archival content, raise from a data protection perspective. These questions are particularly relevant within the EU, where the General Data Protection Regulation ('GDPR') applies.⁶ A deal secured by OpenAI with GEDI, one of Italy's largest publishers, in 2024 has triggered the intervention of the Italian data protection authority ('DPA').⁷ The Italian DPA issued a formal warning to both companies, citing a possible violation of their obligations under the GDPR.⁸ The DPA stresses the particularly vast scale of processing of personal data from the journalistic content of the GEDI titles which would take place under the agreement, as well as an allegedly insufficient level of assessment by the companies of their compliance responsibilities.⁹ At the time of writing, the deal has been temporarily suspended.¹⁰

This article aims to unpack the data protection implications of the increasingly frequent partnerships between AI companies and news publishers.

¹. Michael M Grynbaum and Ryan Mac, 'The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work' *The New York Times* (New York, 27 December 2023) <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html> accessed 10 August 2025.

². *The New York Times Company v Microsoft Corporation*, 1:23-cv-11195, (S.D.N.Y. Dec 27, 2023) ECF no 1, para 2.

³. Ibid; Eduardo Suarez, 'New York Times publisher A.G. Sulzberger: "Our industry needs to think bigger"' *Reuters Institute* (19 February 2024) <https://reutersinstitute.politics.ox.ac.uk/news/new-york-times-publisher-g-sulzberger-our-industry-needs-think-bigger> accessed 10 August 2025.

⁴. Charlotte Tobitt, 'Who's suing AI and who's signing: Major Japan newspaper sues Perplexity, while Gannet signs' *Press Gazette* (12 August 2025) <https://pressgazette.co.uk/platforms/news-publisher-ai-deals-lawsuits-openai-google/> accessed 10 August 2025.

⁵. 'AI Deals and Disputes Tracker' (*Tow Center for Digital Journalism*, 15 December 2025) <https://tow.cjr.org/ai-deals-lawsuits/> accessed 10 August 2025.; Ibid.

⁶. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR) [2016] OJ L119/1, art 58 (2) (a).

⁷. OpenAI, 'OpenAI and GEDI partner for Italian news content' (*Open AI*, 26 September 2024) <https://openai.com/index/gedi/> accessed 10 August 2025.

⁸. 'Provvedimento del 27 novembre 2024 [10077129]' (*Garante per la Protezione dei Dati Personali*, 27 November 2024) <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/10077129> accessed 10 August 2025; Jeric Alvin Cruz, 'The GEDI-OpenAI Case: A Warning About the Processing of Sensitive Data for AI Training' *Media Laws* (12 February 2025) <https://www.medialaws.eu/the-gedi-openai-case-a-warning-about-the-processing-of-sensitive-data-for-ai-training/> accessed 10 August 2025.

⁹. Provvedimento del 27 novembre 2024 [10077129]' (*Garante per la Protezione dei Dati Personali*, 27 November 2024) <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/10077129> accessed 10 August 2025;

¹⁰. The authors have inquired from the Italian DPA about the status of the investigation. GEDI Group, 'Press Release Data Protection Authority' (*GEDI Group*, 29 November 2024) <https://www.gedi.it/en/media/news/gedi-group-press-release-data-protection-authority> accessed 20 July 2026.

1.2 Research Question, outline, and terminology

The content of the licensing agreements, including all relevant GDPR-documentation, remains often undisclosed. Therefore, we cannot exactly determine the type of processing that is taking place in each scenario. It can nevertheless be assumed that the processing activities derive from the set of techniques commonly used to leverage large datasets in LLMs, namely pre-training, fine-tuning and Retrieval Augmented Generation (RAG). Here, numerous questions are worth exploring under the GDPR. Our analysis focuses, however, on the threshold question of whether these deals can rely on a valid legal basis at all. Moreover, because this any further evaluation presupposes knowledge of the existence and scope of these arrangements, we also examine the transparency obligations that may arise for LLM-providers and publishers.

We therefore ask ourselves: ‘How do different choices for technical integration of journalistic content into LLMs in the context of licensing deals between news publishers and AI companies affect compliance with lawfulness and transparency requirements under the GDPR?’. We tackle this question by first detailing the technical solutions available for the integration of journalistic content into LLMs. Secondly, we consider the applicability of the GDPR, assessing how different technical solutions affect the allocation of processing responsibilities and the lawfulness and transparency questions the partnerships face. Moreover, we explore their potential link to the special processing regime for journalistic purposes under art 85 (2) GDPR, which could partially or totally relieve the partnerships from the aforementioned obligations.

The first chapter relies on a review of the relevant technical literature. The remaining parts of the article draw on doctrinal legal research.¹¹ The choice aligns with the purpose of this article of providing a systematic overview of the data protection law, including cases from the Court of Justice of the European Union (CJEU) and Opinions of the European Data Protection Boards (EDPB), applicable to the case of these partnerships.

Private partnerships involving the sale or licensing of personal data for the improvement of AI models and services are certainly no novel issue within the data protection scholarship.¹² Still, we contend that the peculiarities *ratione materiae* and *ratione personae* of these tech-publisher partnerships are worth exploring as a separate issue within the broader field. The personal data at stake is also journalistic information, which the press has a fundamental right to impart and the public to receive.¹³ The partnerships therefore seem to introduce new frontiers to old tensions between media freedom and data protection.¹⁴ Despite the substantial overlap with the discussion unfolding around the non-consensual scraping of online content (including journalistic articles), we still consider these deals to deserve their own spotlight. This is because, their consensual nature alludes to publishers’ explicit assumption of responsibilities under the GDPR. This is important, considering that these partnerships may also institutionalise processing activities inconsistent with the GDPR under the guise of publishers’ freedom to disseminate information. The processing of personal data in the context of LLMs is architecturally at odds with several GDPR principles.¹⁵ Moreover, the epistemic limitations of LLMs can often lead to erroneous reproduction of personal data about individuals through consumer-facing chatbots.¹⁶ This may be more problematic where the press actively participates to the processing, considering its ‘duties and responsibilities’ under European human rights law with respect to the balancing of citizens’ privacy and the provision of ‘reliable and precise information’.¹⁷

^{11.} Sanne Taekema and Wibren van der Burg, ‘Legal Philosophy as an Enrichment of Doctrinal Research Part I: Introducing Three Philosophical Methods’ (2020) *Law and Method* <https://www.boomportal.nl/tijdschrift/LaM/lawandmethod-D-19-00006> accessed 23 March 2026.

^{12.} Janos Meszaros and Chih-hsing Ho, ‘AI research and data protection: Can the same rules apply for commercial and academic research under the GDPR’ (2021) 41 *CLSR*.

^{13.} *Sunday Times v United Kingdom* App no 6538/74 (ECtHR, 26 April 1979), para 65.

^{14.} See for instance, Eric Barendt, ‘Balancing Freedom and Privacy: The Jurisprudence of the Strasbourg Court’ (2009) 1 (1) *J Media L* 49.

^{15.} Hannah Ruschemeier, ‘Generative AI and Data Protection’ (2025) 1 *Cambridge Forum on AI: Law and Governance* 6.

^{16.} James Andrews, ‘What Starbuck v Google Reveals About AI Liability’ *The Regulatory Review* (22 December 2025) <https://www.theregview.org/2025/12/22/andrews-what-starbuck-v-google-reveals-about-ai-liability/>, accessed 24 March 2026; Walter Quattrociochi and others, ‘Epistemological Fault Lines Between Human and Artificial Intelligence’ (*arXiv*, 22 December 2025) <https://arxiv.org/abs/2512.19466> accessed 24 March 2026.

^{17.} *Magyar Helsinki Bizottság v Hungary*, App no 18030/11 (ECtHR, 8 November 2016); *Bladet Tromsø and Stensaas v Norway*, App no 21980/93 (ECtHR, 20 May 1999), para 65.

For the sake of clarity, we use the term ‘LLM-provider’ as a reference to the AI companies (such as OpenAI, Anthropic, Mistral etc.) engaged in the development of LLMs and the operation of consumer-facing chatbot applications like ChatGPT, Claude, LeChat, based on these models. For the sake of simplification, we have also coined the term ‘tech-publisher partnership’ to refer to the licensing deals.

2. LLMs and Journalistic Content 101

In this section, we first outline what technical solutions are available on which these partnerships may rely to enrich the output of an LLM with publishers’ journalistic content. We briefly introduce how each of these solutions works, focusing on the underlying (personal) data processing.

There are several ways to use journalistic content to enhance the capabilities of AI systems based on LLMs. The first and most obvious way is to use the text in the pre-training of the LLM. During pre-training the publishers’ data can be used to influence the parameters of the model and embed the information of the journalistic content into the model. The second technique we describe is model ‘fine-tuning’. This is particularly useful when the model needs to be trained to perform a specific task, such as generating summaries of news articles.¹⁸ The third and most agile way to use news articles we discuss is Retrieval Augmented Generation (RAG). This technique can be applied at the time of generating model output and is less rigid, as it does not involve (re)training the model.¹⁹

2.1 Journalistic content as training data

Current state-of-the-art LLMs are based on a deep learning architecture known as the ‘transformer’.²⁰ This technology uses algorithms to create a model that outputs predictions based on patterns in data. For LLMs these predictions are often the next token in a text.²¹ A token can be a word or a part of a word, the exact length and structure of a token depend on the design of the LLM.²²

Using news articles in the training data can influence the prediction of the next token. It can both affect the capability of the model to reproduce information as well as change the tone or style of the texts produced by the LLM.²³ Reproduction by an LLM can also include, more or less verbatim, sequences of training data when provided with the right contextual input.²⁴ This is called memorisation and should not be seen as a buggy feature, leading to episodic incidents, but rather as a crucial and deliberate characteristic of LLMs’ architecture.²⁵ Despite limited understanding of the phenomenon, current research hints to a strong correlation between memorisation and (increasing) model parameter size and training duration, especially for names of persons or entities and other uncommon sequences.²⁶

¹⁸ Edward J Hu and others, ‘LoRA: Low-Rank Adaptation of Large Language Models’ (*arXiv*, 16 October 2021) 2 <http://arxiv.org/abs/2106.09685> accessed 25 August 2025.

¹⁹ Yunfan Gao and others, ‘Retrieval-Augmented Generation for Large Language Models: A Survey’ (*arXiv*, 27 March 2024) 5 <http://arxiv.org/abs/2312.10997> accessed 20 August 2025.

²⁰ Alec Radford and others, ‘Language Models Are Unsupervised Multitask Learners’ (2019) 4 <https://api.semanticscholar.org/CorpusID:160025533> accessed 20 July 2026.

²¹ Ashish Vaswani and others, ‘Attention Is All You Need’ (*arXiv*, 2 August 2023) 5 <http://arxiv.org/abs/1706.03762> accessed 8 May 2025.

²² Mehdi Ali and others, ‘Tokenizer Choice For LLM Training: Negligible or Crucial?’ in Kevin Duh, Helena Gomez and Steven Bethard (eds), *Findings of the Association for Computational Linguistics: NAACL 2024* (Association for Computational Linguistics 2024), 3907.

²³ Inacio Vieira and others, ‘How Much Data Is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes’ in Rebecca Knowles, Akiko Eriguchi and Shivali Goel (eds), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (Association for Machine Translation in the Americas 2024).

²⁴ Alexander Xion and others, ‘The Landscape of Memorisation in LLMs: Mechanisms, Measurement, and Mitigation’ (*arXiv*, 8 July 2025) 2 <https://arxiv.org/abs/2507.05578> accessed 7 January 2026.

²⁵ Henrik Nolte, Michèle Finck and Kristof Meding, ‘Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data’ (*arXiv*, 3 March 2025) <https://arxiv.org/abs/2503.01630> accessed 7 January 2026.

²⁶ Xion and others (n 24) 3-6.

Changing the output of the model through pre-training is a rigid and resource-intensive process. To significantly alter the patterns learned from the data, the model would have to be essentially recreated from scratch.

2.2 Fine-tuning

Fine-tuning a model also involves using data to influence the parameters of the model.²⁷ In this case the starting point is an already trained LLM which can be adapted for increased accuracy in the performance of a specific task.²⁸ This can rely on techniques such as Low-Rank Adaptation (LoRA), targeting the performance of specific parts of the model, depending on the type of task. For instance, an LLM can be fine-tuned to more accurately summarise news articles. This is done by curating a dataset with the original articles and the target summary.²⁹

Just like for pre-training, the personal data in the fine-tuning dataset is processed in the training process, gets embedded in the model, which is then capable of generating personal data in the output based on the used datasets. The amount of data needed for fine-tuning does however differ significantly compared to pre-training.³⁰

2.3 Retrieval Augmented Generation

After training a model, the information embedded in the model is fixed and will quickly become out of date. A technique to introduce new information without retraining the model is RAG.

RAG involves creating an index of several external documents, based on which an LLM can retrieve relevant information.³¹ A document can, for example, be a news article on a publisher's website, but also any text. What information is retrieved, depends on the prompt the user of the LLM-based chatbot writes. The retrieved information is then augmented to the prompt of the user. This means that additional information is included in the text input to the LLM, instead of just the prompt. This process takes place before inference is ran on the LLM and is independent of the LLM itself.³²

This process steers the generation of output so as to incorporate the information retrieved from the document. There is evidence of increased factual accuracy in model output when relying on this technique.³³ However, there is no absolute guarantee as to the correctness of the output.

Within the RAG pipeline, multiple processing steps that involve personal data can be identified. The first step is the collection of the documents containing personal data. Subsequently, an index needs to be created through a machine learning model processing these documents and creating a mathematical representation of them. In our case scenario, this could be a mathematical representation of the text of news articles contained in journalistic content.

²⁷ Hu and others (n 18) 2.

²⁸ D.M. Anisuzzaman and others, 'Fine-Tuning Large Language Models for Specialized Use Cases' (2025) 3 (1) Mayo Clinic Proceedings: Digital Health 100184, 1.

²⁹ Dongqi Liu and Vera Demberg, 'RST-LoRA: A Discourse-Aware Low-Rank Adaptation for Long Document Abstractive Summarization' in Kevin Duh, Helena Gomez and Steven Bethard (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (Association for Computational Linguistics 2024), 2204.

³⁰ Inacio Vieira and others, 'How Much Data Is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes' in Rebecca Knowles, Akiko Eriguchi and Shivali Goel (eds), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (Association for Machine Translation in the Americas 2024).

³¹ Patrick Lewis and others, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', *Advances in Neural Information Processing Systems* (Curran Associates, Inc 2021); Uday Kamath and others, 'Retrieval-Augmented Generation' in Uday Kamath and others (eds), *Large Language Models: A Deep Dive: Bridging Theory and Practice* (Springer Nature Switzerland 2024), 276.

³² This means that the augmentation of the prompt of the user happens before the LLM is used to generate an answer.

³³ Lewis and others (n 31), 6.

After creating an index, every time the documents are queried to retrieve relevant information, the information in the index is processed. The retrieved information varies based on its relevance to the prompt of the user. The output of this process, at this stage, can include personal information, depending on the journalistic content. This data is then further processed during augmentation and, ultimately, in the LLM's generation of output.

It is not clear from press releases and other publicly available documentation, which technical solutions the tech-publisher partnerships pursue. We suspect that most primarily involve pre-training or fine-tuning, while RAG may be applied where these announcements mention up-to-date news distribution or output generation referencing links to publishers' articles.³⁴ Therefore, the legal analysis that follows is applied to all three described solutions, treating pre-training and fine-tuning as a single category due to their functional similarity (model development).

In the next chapter, we introduce the GDPR and its applicability to these partnerships. Furthermore, we reflect on the extent to which these partnerships may qualify for the special processing regime for journalistic purposes under art 85 (2) GDPR.

3. Licensing journalistic content under the GDPR

3.1 Material, personal, and geographic scope

The applicability of the GDPR to tech-publisher partnerships depends on whether the licensed journalistic content contains personal data, namely information about identified or identifiable subjects.³⁵ Non-identifiable data does not fall within the scope of the GDPR.³⁶

The data contained in journalistic content is basically the totality of information published by direct employees of the publisher and freelancers in their professional activity. Despite journalistic ethical codes requiring reporters to exercise utmost care when deciding how to publicly disclose information concerning individuals, this ambition often does not translate easily into practice. Uncovering information is after all the essence of the journalistic profession, which may often clash with the privacy of individual subjects. Invasive reporting is usually linked to the commercial incentive media outlets face to publish sensationalistic stories. On the other hand, societal shifts toward openness are increasingly blurring the boundaries of public and private life. Medical conditions, criminal convictions, sexuality, and other traits of personal histories are now essential components of broader political debates, rendering the journalist's task of balancing public interest with respect for individual rights to privacy and data protection ever more challenging.³⁷ Despite the fact that much of the reporting done in news media concerns the activities of politicians, entertainment celebrities, and corporate stars – who in any case still enjoy a fundamental right to privacy –³⁸ detailed accounts of ordinary individuals may often come to surface through in-depth interviews and local reporting. In time, news archives have therefore also developed into vast repositories of personal, as well as sensitive, information protected under the GDPR.

When news organisations license journalistic content to LLM-providers for either training, fine-tuning or as a source for RAG, the GDPR applies. Due to the territorial scope defined in art 3 GDPR, however, this is true only for deals in which (a) either party is established in the EU, (b) the personal data contained in journalistic content relate to EU residents and either party offers goods or services to individuals in the EU.³⁹ Any deal between AI companies and European publishers is likely to satisfy these conditions. In theory, the extra-territorial scope of the GDPR could stretch even to partnerships concluded outside the EU.

³⁴ See for instance, Reuters, 'Meta Platforms to use Reuters news content in AI chatbot' *Reuters* (25 October 2024) <https://www.reuters.com/technology/artificial-intelligence/meta-platforms-use-reuters-news-content-ai-chatbot-2024-10-25/> accessed 7 January 2026.

³⁵ GDPR, art 4 (1).

³⁶ GDPR, recital 26.

³⁷ Anita L Allen, 'Why Journalists Can't Protect Privacy', in Craig LaMay (ed) *Journalism and the Debate over Privacy* (Routledge 2003), 73.

³⁸ See *Von Hannover v Germany* App no. 59320/00 (ECtHR, 24 June 2004).

³⁹ GDPR, art 3.

In charting the different implications raised from a GDPR perspective by tech-publisher partnerships, we take into account the three main technical arrangements for the integration of journalistic content into LLMs described above (model training/fine-tuning and augmentation through RAG). We first seek to understand how these modalities of processing translate into different responsibilities under the GDPR. Based on this outcome, we proceed to discuss several substantive obligations, including the need for a valid legal basis and transparency of processing.

3.2 GDPR Responsibilities

The first substantive question that needs to be tackled is what processing responsibilities the parties involved in the partnerships assume under the GDPR.

The GDPR allocates data protection responsibilities across two main categories of natural, legal persons and public bodies involved in the processing: controllers and processors. The former are defined as those who determine the ‘purposes and means’ of personal data processing, whether acting individually or jointly.⁴⁰ The GDPR envisages the possibility of ‘joint’ controllership, where two or more controllers contribute to the determination of purposes and means.⁴¹ Processors instead merely process personal data on behalf of the controller(s).⁴²

The distinction is crucial because controllers remain legally responsible for implementing ‘appropriate technical and organisational measures’ to demonstrate compliance, even with respect to the activities conducted by the contracted processor(s).⁴³ Processors are only responsible for their own actions in complying with the instructions provided by the controller(s) and other specific obligations under national and EU law (including the GDPR).⁴⁴ Failure to properly identify the relevant responsibilities may therefore result in an inadequate distribution of responsibilities and consequent loopholes in the protection of data subjects.

The EU authorities’ discussion of data processing responsibilities, and in particular controllership, has a long-standing history, which has duly been detailed and criticised in the literature.⁴⁵

3.2.1 (Joint) Controllers?

Determining the GDPR responsibilities in the tech-publisher partnerships requires us to take a closer look at how the contractual parties are *factually* positioned to determine the purpose and means of processing.⁴⁶ Terms of a contractual agreement, such as a data processing agreement, may play a role in the analysis, but are not ultimately decisive if they are not reflective of factual circumstances.⁴⁷

Whether for training, fine-tuning or RAG, the processing chain consists of three main stages: data access, data curation, and data integration in the model. Making the data available for the LLM-provider is, of course, a prerequisite for the execution of the deals. This can be, for example, accomplished by transferring the dataset as a whole or providing access to ‘live’ data through an API. Curation involves selecting the data point that should be used for RAG, training or fine-tuning. The purposes of transfer and curation should

⁴⁰ GDPR, art 4 (7).

⁴¹ GDPR, art 26 (1).

⁴² GDPR art 4 (8).

⁴³ GDPR, art 82 (2).

⁴⁴ GDPR, art 82 (2).

⁴⁵ See Lee A Bygrave and Luca Tosoni, ‘Article 4(7). Controller’ in Christopher Kuner and others (eds) *The EU General Data Protection Regulation (GDPR)* (Oxford University Press 2020); See van Mill and Quintais, *A Matter of (Joint) control? Virtual assistants and the general data protection regulation*, (2022) *Computer Law & Security Review*, 15; René Mathieu and Joris van Hoboken, ‘Fashion-ID: Introducing a phase-oriented approach to data protection?’ (*European Law Blog*, 30 September 2019) 5 <https://www.europeanlawblog.eu/pub/fashion-id-introducing-a-phase-oriented-approach-to-data-protection/release/1> accessed 7 January 2026.; Christopher Millard and others, ‘At this rate, everyone will be a [joint] controller of personal data’ (2019) 9 (4) *International Data Privacy Law* 217; Michèle Finck, ‘Cobwebs of control: the two imaginations of the data controller in EU law’ (2021) 11 (4) *International Data Privacy Law* 333.

⁴⁶ This is unless national or EU law explicitly designates data processing responsibilities. GDPR, art 4 (7). See ‘Guidelines 07/2020 on the concepts of controller and processor in the GDPR’ (*European Data Protection Board (EDPB)*, 7 July 2021) 11-12.

⁴⁷ EDPB (n 46)

not be assessed in isolation, because they are ultimately dependent on the subsequent use of the data in the LLM.

Publishers as controllers

Journalists collect personal data in an article, which the publisher then stores, disseminates and subsequently makes available for the LLM-provider. Until this moment, the publisher is in full control of the purpose and means of the processing. Hence, the publisher acts as a controller at this stage, whether the data is transferred for model training and fine-tuning or as an external source for RAG.

LLM-providers as controllers

Upon gaining access to the data, we find that LLM-providers should also be considered controllers. If publishers make their data available for training or fine-tuning, providers have indeed exclusive discretion in determining the why and how of processing, thus acting as controllers.

Things differ a bit where they only gain access to journalistic content for RAG. One could argue that LLM-providers could be seen in this case as processors disseminating news on behalf of the publisher – by curating the data as context for the chatbot's generation of responses. Nevertheless, we would argue that, also here, there is a case for controllership. From the CJEU's ruling in *Google Spain*, we understand that search engine operators contributing to the dissemination of news articles published online should also qualify as controllers.⁴⁸ The CJEU saw Google's indexation of the news article and publication of the relevant hyper-links as distinct and additional processing activities from those of the publishers of which only Google could determine purposes and means.⁴⁹ LLMs and (traditional) search engines belong to distinct technological categories, but, as the Dutch DPA also acknowledges, they share several functional overlaps between them, including their contribution to the dissemination of digital information.⁵⁰ This is also consistent with the CJEU's expansion of controllership assignment in a diverse range of case scenarios for the sake of ensuring 'effective and complete protection of data subjects', albeit through a controversial dilution of the qualifying threshold.⁵¹

That publishers retain control over what information they publish and that they may profit from innovative news dissemination channels does not take away the fact that LLM-providers still exert the largest share of influence over purpose and means of execution. It is in fact the LLM-provider deciding on the parameters governing which relevant articles are selected and how prompts are augmented.

Joint controllership?

LLM-providers and news publishers are therefore controllers, regardless of the type of processing involved in the partnerships. What could change, though, is the nature of the relationship between the two controllers, namely whether different processing activities could give rise to joint or separate controllership.

The existence of joint controllership can have important legal consequences. The GDPR expects joint controllers to agree on the allocation of responsibilities to ensure compliance with the rest of the regulation and make the 'essence' of this understanding available for data subjects.⁵² Moreover, joint controllers remain jointly liable for the entire damage arising from a GDPR violation across the processing value chain.⁵³ The responsibility of separate controllers is instead confined to the scope of their individual controllership.⁵⁴

⁴⁸ Case C-131/12 *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* [2014] ECLI:EU:C:2014:317, para 33.

⁴⁹ *Google Spain* paras 35-36.

⁵⁰ Verantwoord vooruit: AVG-Randvoorwaarden voor generatieve AI' (Autoriteit Persoonsgegevens, 23 May 2025), 12.

⁵¹ See *Google Spain*, para 34; Case C-40/17 *Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW eV* [2019] ECLI:EU:C:2019:629; Case C-210/16 *Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH* [2018] ECLI:EU:C:2018:388; C-25/17 *Jehovan todistajat* [2018] ECLI:EU:C:2018:551; Michele Finck even speaks of 'pulverisation' of control. See Finck (n 45) 335-337.

⁵² GDPR, art 26 (1)-(2).

⁵³ GDPR, art 82 (4). The CJEU however also stresses that the extent of liability should be established for each party on the basis of all the relevant circumstances, including their involvement at different stages of processing. *Jehovan todistajat*, para 66.

⁵⁴ This is subject to special civil liability provisions under national law.

Where journalistic content is transferred for the purpose of model training or fine-tuning, we contend that this is a case for separate controllership. The LLM-provider uses the data of the publisher for a purpose on which the publisher bears no influence or might not even be aware of. What happens after transfer of the data is not relevant for publishers' GDPR compliance, provided that they fulfil all obligations applicable to their own processing activities.⁵⁵

Joint controllership instead arises where journalistic content is used as a source for RAG. Several controllers may be involved 'at different stages' and 'to different degrees', without necessarily bearing the same level of responsibility for each of those.⁵⁶ This, however, does not exclude the establishment of joint controllership for those processing stages in which controllers converge on purposes and means.⁵⁷ To use the terminology of the EDPB, this is a situation in which converging decisions on the purposes and means made by the different parties should be seen at the 'macro-level' as complementary to each other and necessary for the processing to take place.⁵⁸ This means that different processing operations by different parties are 'inextricably linked' and cannot be considered as separate if the processing as a whole pursues a joint purpose through 'jointly defined means'.⁵⁹

In the case of RAG, there is a converging purpose, or at least two purposes which are 'closely linked' or 'complementary',⁶⁰ namely news dissemination/ enriching model output with journalistic information. Secondly, the fact that LLM-providers exert legal and technical control over most of the processing infrastructure does not necessarily imply that news organisations lack sufficient influence over the means of processing.⁶¹ Provided that an LLM-provider refrains from using the journalistic content for purposes outside the jointly agreed objective, the relationship qualifies as joint controllership despite its predominant influence over the processing infrastructure.

In the next section, we show how the discussed technical solutions and corresponding processing responsibilities relate to the fulfilment of the lawfulness and transparency principles under art 5 (1) (a) GDPR.⁶²

3.3 Can these deals rely on a legal basis under art 6 (1) GDPR?

Lawfulness is the first of the nine principles which controllers and processors need to take into account when engaging in the processing of personal data. Abiding by the principle of lawfulness requires that processing respects all legal requirements that may be applicable to it (even beyond the GDPR).⁶³ According to art 6 (1) GDPR, personal data processing must be grounded in one of the legal bases listed therein. Out of the six alternatives listed, we consider only (a) individual consent and (f) legitimate interests of the controller or a third party as potentially applicable legal grounds to the processing operations envisaged in the licensing deals. Art 6 (1) (b), allowing processing where this is necessary for the performance of a contract to which the data subject is a party, cannot apply due to the data subjects not being involved as parties in the licensing deals. The same counts for art 6 (1) (c, d, and e), which are totally out of scope. It should be borne in mind that all controllers involved in (joint) processing operations must be able to rely on a valid legal basis.⁶⁴

⁵⁵ See *Fashion ID*, para 74.

⁵⁶ *Wirtschaftsakademie*, para 43; *Fashion ID*, para 76.

⁵⁷ *Wirtschaftsakademie*, para 43; *Fashion ID*, para 76; Cfr with Case C-604/22 *IAB Europe v Gegevensbeschermingsautoriteit* [2024] ECLI:EU:C:2024:214, paras 74-76.

⁵⁸ EDPB (n 46) 19.

⁵⁹ EDPB (n 46) 17-20.

⁶⁰ EDPB (n 46) 17-20.

⁶¹ EDPB (n 46) 21; See *Wirtschaftsakademie*, paras 35-39.

⁶² GDPR, art 5(1) (a).

⁶³ Cecile de Terwangne, 'Article 5. Principles relating to processing of personal data' in *The EU General Data Protection Regulation (GDPR)* (2020), 314.

⁶⁴ EDPB (n 46) 45.

In a 2024 Opinion, the EDPB also deals with the question of the appropriate legal basis under the GDPR for model training (and fine-tuning) separately from deployment, the latter including the use of RAG for output augmentation.⁶⁵

3.3.1 Consent as a legal basis

The first potentially valid legal basis to process personal data in the context of tech-publisher partnerships is obtaining informed consent from data subjects, whose personal data is contained in the shared journalistic content. According to the GDPR, consent must be ‘freely given, specific, informed and unambiguous’.⁶⁶

The choice of consent as a legal basis for either training or RAG is clearly impractical for both controllers due to the large amount of data subjects involved and, for already published articles, the difficulty of tracing them even across several countries. Moreover, a consenting data subject maintains the right to withdraw previously given consent at any moment.⁶⁷ This could raise further practical hurdles, besides limiting the pool of data from which LLMs would draw.⁶⁸

Most reporting occurs anyway regardless of data subjects’ consent in light of journalistic freedom protected under fundamental rights law and even the GDPR.⁶⁹ It is then for DPA’s and courts to assess whether that freedom is properly exercised vis-à-vis the fundamental right to privacy of the data subjects.

The only situation in which consent might be relevant is journalists’ routine requests for sources to participate in interviews or release statements for publication. Here, we could legitimately consider consent as a practical solution to address the need for a legal basis for the execution of tech-publisher partnerships. For instance, Bruce Gillespie advocates for a quality upgrade of sources’ consent to strengthen (vulnerable) interviewees’ awareness of the increasing information retrieval capabilities in the digital environment in which their data is published.⁷⁰ In case of RAG, the joint controllers could formalise this step in their joint controllership agreement to justify processing.⁷¹ Instead, if the partnership involves model training, the scope of processing activities legitimised by this consent would be limited to those activities over which publishers act as controllers, namely up to the data transfer. As separate controllers, LLM-providers would therefore need to obtain this separately for their own purposes.

3.3.2 Legitimate Interest as a legal basis

The limits associated with the obtainment of consent make the pursuit of legitimate interest(s) under art 6 (1) (f) GDPR a more appetible legal basis for these partnerships. It is indeed not surprising that GEDI and OpenAI relied on this ground, as reported by the Italian DPA.⁷² The existence of a legitimate interest is assessed on the basis of a tripartite test.⁷³ Again, depending on the actor and the type of processing activity, our analysis can lead to different implications.

First Step – The existence of a legitimate Interest

First, the controllers must be able to invoke a legitimate interest for processing. Despite the CJEU’s emphasis on the need for a ‘restrictive’ application of the art 6 (1) (f),⁷⁴ its understanding of ‘interest’ is quite flexible,

⁶⁵ ‘Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models’ (EDPB, 18 December 2024) 2.

⁶⁶ GDPR, art 4 (11).

⁶⁷ GDPR, art 7 (3).

⁶⁸ An alternative option would be further processing personal data collected on a valid legal basis for purposes compatible with the original one, under art 5 (1) (b) GDPR. The issue with relying on purpose compatibility is that, even if it were possible, it would only count for data collected on the basis of a legal basis under art 6 (1) GDPR, like consent.

⁶⁹ See discussion below on the special processing regime under the GDPR for journalistic purposes.

⁷⁰ See Bruce Gillespie, ‘The Case for Using Informed Consent in Journalism’ in Lada Trifonova Price and others (eds) *The Routledge Companion to Journalism Ethics* (Routledge 2021).

⁷¹ GDPR, art 26 (1).

⁷² Garante per la Protezione dei Dati Personali (n 8).

⁷³ Case C-13/16 *Valsts policijas Rīgas reģiona pārvaldes Kartības policijas pārvalde v Rīgas satiksme* [2017] ECLI:EU:C:2017:336, para 28.

⁷⁴ Case C-252/21, *Meta Platforms Inc and Others v Bundeskartellamt* [2023] ECLI:EU:C:2023:537, para 93.

as long as this is effective and clear and not too generic.⁷⁵ In recent case law, it could be successfully invoked even for product improvement, as well as pure commercial interests.⁷⁶

From the perspective of an LLM-provider, it is therefore likely that art 6 (1) (f) GDPR could serve as a legal basis for the improvement of its model or other product (i.e. online chatbot) through training or fine-tuning. Improvement could be interpreted narrowly, referring specifically to the model's ability to disseminate accurate journalistic content, or more broadly, referring to the enhancement of its overall performance.

The EDPB is positive on the suitability of the legal basis for model training, depending on a case-by-case assessment. Some national DPA's take clearer stances. The Spanish DPA, for instance, states that legitimate interests could justify processing requiring access to training data.⁷⁷ The French CNIL shares comparable views, even where the processing serves general improvement of the model's functioning, though the precise purpose of deployment might be unknown at the time.⁷⁸ The Dutch DPA sees legitimate interest as the 'only possible basis' for the training of foundation models, provided it meets other data protection requirements.⁷⁹ In line with increasingly lenient case law of the CJEU, improving one's model by feeding it high quality journalistic content should pass the first step of the tripartite test under art 6 (1) (f) GDPR. The same conclusion, by analogy, is valid when the model processes the data through RAG in the deployment phase.

From the perspective of publishers, the making available of journalistic content for LLM-providers to train or fine-tune their models on could find a legitimate interest in their pursuit of commercial profit (through licensing fees) and/or dissemination of journalistic content. Alternatively, assuming a model provider could prove its own legitimate interest, a publisher has the option under art 6 (1) (f) to rely on its partner's interest. In any case, the legal basis should provide a valid ground for processing up to the moment the data transfer between the two contractual parties takes place. Whatever happens afterwards should be within the domain of controllership of the LLM-provider acting as a sole controller. It is interesting to note that one of the grounds on which the Italian DPA urged GEDI and OpenAI to revisit their plans was the inadequacy of art 6 (1) (f) GDPR to justify the transfer of GEDI's journalistic content to OpenAI. It should be pointed out, however, that GEDI did not explicitly mention its commercial interests, opting instead for its interest in pursuing 'journalistic activities' through innovative news dissemination. In any case, the Italian DPA does not see how GEDI's own interests could justify data processing for the purpose of model training.⁸⁰ Should this opinion stir consensus across the EU, publishers could try to rely on the aforementioned third-party interest (of the LLM-provider), though this also offers limited room.⁸¹

When journalistic content is used as a source for RAG, the publisher remains jointly responsible as co-controller with the LLM-provider throughout the processing chain. As for pre-training and fine-tuning, each party would require its own legal basis, which could well be legitimate interest. However, as per CJEU case law, they would not be able to rely on the other's legitimate interest, due to being joint controllers, but would need to prove this individually.⁸² Again, the LLM-provider could rely on product improvement goals, in particular providing users with accurate and up to date information on most recent news. The partner publisher should be able to rely on the journalistic innovation argument more comfortably than in the

⁷⁵ See GDPR, recital 47; Waltraut Kotschy, 'Article 6. Lawfulness of processing' in Christopher Kuner and others (eds) *The EU General Data Protection Regulation (GDPR)* (Oxford University Press 2020); Case C-708/18 *Asociația de Proprietari bloc M5A-Scara A* [2020] ECLI:EU:C:2019:1064, para 44; 'Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR' (EDPB, 20 November 2024) 9.

⁷⁶ CJEU *Meta Platforms*, paras 122-123; Case C-621/22 *Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens* [2024] ECLI:EU:C:2024:858.

⁷⁷ Agencia Española Protección Datos, 'Adequación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción' (2020) 22.

⁷⁸ CNIL, 'IA Mobiliser la base légale de l'intérêt légitime pour développer un système d'IA' (2025) <https://www.cnil.fr/fr/base-legale-interet-legitime-developpement-systeme> accessed 1 October 2025.

⁷⁹ Autoriteit Persoonsgegevens (n 50) 10.

⁸⁰ Garante per la Protezione Dati (n 8).

⁸¹ Kotschy (n 75) 337.

⁸² *Fashion ID*, para 96.

case of model development. The only caveat here is that, as a joint controller, the publisher could be held responsible across the processing chain, rather than just up to the moment of transfer, possibly raising its compliance burden.

We discuss the remaining steps of the test jointly.

Second and Third Steps – Necessity and Proportionality

At the second and third stages, it should be proven that the processing of journalistic content is necessary to achieve the above discussed interests and that the interests of data subjects affected are not overriding. In the context of art 6 (1) (f), the CJEU has explained that necessity entails determining whether the legitimate interest could not be reasonably pursued just as effectively through other means bearing less on the fundamental rights of the data subjects.⁸³ This closely aligns with the objective of the data minimisation principle, to which controllers are bound under art 5 (1) (c) GDPR.⁸⁴ The third step instead essentially boils down to a proportionality test *stricto sensu* – the more classical ‘balancing exercise’ aimed at determining whether the pursuit of the controllers’ interests over those of the data subjects is justified at all.⁸⁵

When it comes to model pre-training or fine-tuning, we think that both LLM-providers and publishers could have a hard time at these stages. This is due to the aforementioned collateral effect of memorisation in the training process. Feeding journalistic content as training datasets to LLMs implies permanent embedding of the personal data contained therein in the model’s parameters, which can lead to verbatim reproductions during inference.

When assessing the second limb of the test – necessity of processing – model providers should not be able to justify training operations for product improvement in the narrow sense (accurate dissemination of journalistic content). As argued below, this is suitably done through RAG without the need for personal data memorisation. They could argue, however, that pre-training or fine-tuning would be the most effective method to improve the model’s general performance, considering that alternative approaches like RAG do not incorporate new information into the model’s learned representations.

Still, the claimed legitimate interest might not hold ground at the third limb of the test due the undesirable effects of memorisation on data subjects’ rights. The GDPR prescribes several criteria relevant to the assessment, including the nature of the data, the controller-data subject relationship, the reasonable expectations of the data subjects, the further consequences of processing, and the possible mitigating measures controllers could take.⁸⁶

Permanent memorisation in the model of sensitive information of even not-so-public figures contained in journalistic content could also entail permanent availability of that information across different contexts in which the model is deployed, increasing the possibilities of information inference and profiling. This is also considering that older model releases are kept available as default options for users, even after new versions become available for premium subscribers. In such cases, information memorised during training may not be updated over time, meaning that subsequent corrections, updates, or rectifications made by journalists may not be reflected in these models. Moreover, although we cannot claim with certainty that new model versions are trained on the basis of their predecessors, given the scarcity of high-quality data, it is reasonable to assume the AI industry’s incentive in re-using content acquired for the training of a specific release when developing subsequent ones.

It is also not so self-evident how data subjects could expect their data to be licensed for model training, given that this enables an undetermined scope of subsequent processing distant from the original context of collection. This is also considering the lack of a direct relationship between the data subject and the LLM-

⁸³ *Meta Platforms*, para 108.

⁸⁴ *Ibid*, para 109.

⁸⁵ Lorenzo Dalla Corte, ‘On proportionality in the data protection jurisprudence of the CJEU’ (2022) 12 (4) *International Data Privacy Law* 259, 265; EDPB (n 75) 12-13.

⁸⁶ See GDPR, recital 47; EDPB (n 75) 12-13.

provider (or even with the publisher, in case of non-consensual reporting).⁸⁷ The EDPB also suggests that gaining access to personal data through third parties (i.e. through a publisher) rather than through direct disclosure by the data subjects may weaken the presumption that the latter had a reasonable expectation of processing for training purposes.⁸⁸ In the context of social media platforms, the CJEU has already recognized that the impact of mass-scale data processing and the absence of reasonable user expectations pose significant barriers to establishing that third-party data sharing is necessary for product improvement.⁸⁹

Lastly, it is not clear how mitigating measures could tilt the balance in favour of processing. The EDPB mentions data masking and pseudonymisation as relevant safeguards in the model development stage, as well as the facilitation of individual rights exercise.⁹⁰ In the meaning ascribed to it under the GDPR, pseudonymisation could be of no use here. Given the unprecedented amount of data fed into these models, the risk of re-identification during inference is significant.⁹¹ The CJEU's threshold for re-identifiability is remarkably low and would likely be met in such a case.⁹² Once fed to the model, there is also no guarantee that providers would be able to prove compliance with objection to processing or data deletion requests, despite progress on the front of 'unlearning' techniques.⁹³

When relying on RAG to integrate journalistic content, the legitimate interest test's barriers are eased. This is because RAG does not require permanent memorisation of personal data into the model – the data is stored with the retriever only temporarily to provide context during inference.⁹⁴ The impermanent and restricted nature of processing strengthens compliance with the data minimisation principle, while still securing the controllers' interests. Where the relevant interests concern improving a chatbot's journalistic distribution of journalistic content and, as in the case of GEDI, fostering innovation in journalistic processes, RAG can offer an equally effective but less intrusive solution. This undermines the necessity of feeding journalistic content to the model for the (alleged) same purpose.

Furthermore, the impact of processing through RAG on the expectations of data subjects and on their fundamental rights is arguably different than in the case of training. There is no doubt that the information-retrieval capabilities of (chatbots built on) LLMs, offered across language barriers to hundreds of millions of users across the world at the mere prompt of a query, have greater impact than the publication of an article restricted to volatile print or behind a digital paywall.⁹⁵ However, the circumstances that arise with RAG should not be legally different from the processing of journalistic content through search engine requests. Despite recognising a right to be delisted from search engine indexation,⁹⁶ the CJEU has implicitly accepted that data subjects, whose information is published in a news article online can expect that article to be indexed and made available to any search engine users.⁹⁷ Similarly, it would not be far-reaching to say that they could expect that information to be inferred through Mistral's Le Chat, as long as this is retrieved from another consensual source (i.e. publishers).

Compared to model training, the scale and ease of dissemination might not differ with RAG, but at least the information would not be permanently embedded in the model, where update, removal and verification is extremely difficult (though to a lesser extent for fine-tuning). Publishers, as owners and administrators of

⁸⁷ This last is one of the criteria listed in Recital 47 GDPR to conduct the third step of the legitimate interest step and is also stressed by the EDPB in its Opinion. 'Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models' (EDPB, 17 December 2024), 27.

⁸⁸ Ibid.

⁸⁹ *Meta v Bundeskartellamt*, paras 122-123

⁹⁰ EDPB (n 87) 28-29.1

⁹¹ See Xion and others (n 24).

⁹² Case C-582/14 *Breyer v Bundesrepublik Deutschland* [2016] ECLI:EU:C:2016:779, paras 40-49.

⁹³ Xiaoyu Xu and others, 'Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs' (*arXiv*, 22 May 2025) <https://arxiv.org/abs/2505.16831> accessed 7 January 2026.

⁹⁴ See Lewis and others (n 31).

⁹⁵ See a similar argument with respect to scraping of online information of data subjects Taner Kuru, 'Lawfulness of the mass processing of publicly accessible online data to train large language models' (2024) 14 (4) *International Data Privacy Law* 326.

⁹⁶ Case C-136/7 *GC and others v CNIL* [2019] ECLI:EU:C:2019:773, paras 36-37.

⁹⁷ This is implicitly derived from *Google Spain*.

their journalistic content, would still be in control of the source of information across time which, if integrated through RAG in model responses, could mitigate some negative consequences of processing, ranging from improved inference accuracy to more effective data subject redress.⁹⁸

Hence, if controllers decide to rely on legitimate interest as the basis for processing, we find that this could only support the integration of journalistic content through RAG, rather than feeding them during the pre-training and fine-tuning stages.

3.3.3 Exemptions to the processing prohibitions of arts 9 and 10 GDPR.

The last hurdle that controllers would need to overcome is finding a suitable exemption on the prohibition to process data falling under the so-called ‘special categories’ of the GDPR. Art 9 (1) GDPR prohibits any processing of several special categories of personal data, including information revealing racial or ethnic origin, religious and sexual orientation. Processing of this data would only be allowed where controllers can find an additional legal ground under art 9 (2), as well as standard legal basis under art 6 (1).⁹⁹ The issue for tech-publisher partnerships is that no comparable exemption to legitimate interest is listed in art 9 (2) GDPR. Similarly, art 10 GDPR prohibits the processing of personal data relating to criminal convictions and offences if not authorised by EU or national law. The Italian DPA also remarked these points in its warning to GEDI and OpenAI.¹⁰⁰

Whether opting for model training or RAG, the available options to process sensitive data are scarce. For the sake of conciseness, we discuss these only with respect to art 9 (2) GDPR, considering that national implementations of art 10 GDPR generally follow similar approaches.¹⁰¹ The first one is obtaining explicit consent under 9 (2) (a) GDPR. Having already analysed the issues linked to consent as a legal basis under art 6 (1) (a) GDPR and given the more stringent conditions applied under art 9 (2) (a), we can dismiss this.

Secondly, processing sensitive data manifestly made public by the data subject is allowed under art 9 (2) (e). It is quite self-evident that information published consensually by a mass media outlet should meet this definition,¹⁰² meaning that tech-publisher partnerships could in principle rely on this ground, even where these involve model pre-training or fine-tuning on journalistic content. At least, this seems to follow from the CJEU ruling in *Meta v Schrems*, which concerned the processing of data disclosed by the data subject about his sexual preferences during a press conference for targeted advertising purposes by a social media platform.¹⁰³ Still, this does not solve the issue of processing the vast amount of sensitive journalistic data that was not voluntarily manifested by data subjects and included in journalistic content.

Finally, under art 9 (2) (g), there is the option of processing sensitive data for reasons of substantial public interest, based on Union or Member States’ law. A literal interpretation of the provision, and its corresponding recital, leads to the presumption of a high threshold to invoke this special processing ground.¹⁰⁴ However, the CJEU curiously invoked it when stressing a search engine’s function in enabling the exercise of the right to

⁹⁸ See for instance Lewis and others (n 31); Jiarui Li and others, ‘Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases’ (*arXiv*, 15 March 2024) <https://arxiv.org/abs/2403.10446> accessed 7 January 2026; Of course, the technique comes with limitations, Rajendra Gangavarapu and others ‘Evaluating Accuracy in Large Language Models: Benchmarking Corrective RAG vs Naive Retrieval Augmented Generation Approach’ (2025) IEEE ICAD <https://ieeexplore-ieee-org.proxy.uba.uva.nl/document/11114027> accessed 7 January 2026.

⁹⁹ ‘Guidelines 3/2019 on processing of personal data through video devices’ (EDPB, 30 January 2020) 17

¹⁰⁰ Garante per la Protezione Dati (n 8).

¹⁰¹ See for instance the legal basis under the Dutch implementation of art 10 GDPR, Uitvoeringswet Algemene Verordening Gegevensbescherming, art 32 (a).

¹⁰² Ludmila Georgieva and Christopher Kuner, ‘Article 9. Processing of special categories of personal data’ (2020) in Christopher Kuner and others (eds) *The EU General Data Protection Regulation (GDPR)* (Oxford University Press 2020), 378.

¹⁰³ In his Opinion anticipating the *Meta v Schrems* judgment, AG Rantos argued that an online platform could not rely on the publicity of disclosures by a data subject about his sensitive data to process that data for delivering targeted advertising. While conceding that art 9 (2) prevents blanket processing of all other personal data in that particular category, the CJEU did not exclude processing of the specific piece of information made public by the data subject under art 9 (2) (e). See Opinion of AG Rantos, Case C-446/21 *Maximilian Schrems v Meta Platforms Ireland Limited* [2024] ECLI:EU:C:2024:366, para 48; Case C-446/21 *Maximilian Schrems v Meta Platforms Ireland Limited* [2024] ECLI:EU:C:2024:366 .

¹⁰⁴ GDPR, recital 52.

freedom of information of Internet users protected under art 11 of the EU CFR.¹⁰⁵ This should be seen as a pragmatic solution to avoid putting search engine operators in an impossible position and hence disrupting the functioning of the Internet as a whole. Taren Kuru points out that this ‘*de facto* exception’ could provide a springboard for LLM-providers to surmount the roadblock of art 9 (1) GDPR for the processing of sensitive data.¹⁰⁶ The Dutch DPA also does not exclude this possibility in the future.¹⁰⁷ Even if we admit the eligibility of this legal basis for tech-publisher partnerships, we should not forget that art 9 (2) (g) also requires controllers to secure ‘the essence’ of the right to data protection, as well as suitable and specific safeguards for data subjects. Drawing from the discussion on legitimate interests, we can conclude that feeding journalistic content as training data to the models is more unlikely to meet these criteria than using it as sources for RAG.

3.3.4 The EU Commission’s Digital Omnibus Proposal

At the inception of this article, the first rumours about a possible GDPR reform began to spread, yet with nothing concrete in sight. A few weeks after a clumsy public leak, the European Commission officially published the ‘Digital Omnibus’ Regulation on 19 November 2025.¹⁰⁸ The Digital Omnibus contains a series of amendments to the overarching European data law framework, including the GDPR, due to its staggering of the EU’s global competitiveness highlighted in the Draghi and Letta reports.¹⁰⁹

Relevant in the context of this article is the attempt the EU Commission makes at ‘clarifying certain aspects as to the processing of data for AI training and development’.¹¹⁰ This includes an explicit green light for the use of legitimate interest(s) as a legal basis under art 6 (1) (f) GDPR for the processing of personal data in the context of the development and operation of AI models and systems.¹¹¹ An ad hoc derogation from the prohibition to process special categories of personal data is also proposed under art 9 (2) GDPR.¹¹² Both provisions subject processing to the requirement of ‘necessity’ and the application of safeguard measures for the rights and freedoms of data subjects. However, the amendments have been criticised for effectively watering down the protection of data subjects’ control over their own personal data against an increasingly information-hungry AI industry.¹¹³

Assuming the reform could pass as it is, tech-publisher partnerships would have an explicit basis to claim a legitimate interest in the processing of journalistic content as training data for models, including for the special categories of data therein contained. Nevertheless, given that the proposal in its current form does not alter the remainder of the legal test for the assessment of a legitimate interest – namely the necessity and proportionality steps – we think that our analysis above should marginally be affected by the potential amendments, unless the CJEU and the DPAs should interpret these in light of the EU Commission’s deregulatory agenda. The EDPB and EDPS explicitly confirmed this interpretation in their Joint Opinion on the Digital Omnibus submitted to the EU legislator.¹¹⁴

3.4 How transparent should the controllers be?

We come to the last part of the research question, namely the required level of transparency around the tech-publisher partnerships. Transparency is a core GDPR principle requiring accessible and understandable

¹⁰⁵ Case C-136/7 *GC and others v CNIL* [2019] ECLI:EU:C:2019:773, paras 61-66.

¹⁰⁶ Kuru (n 95) 346.

¹⁰⁷ Autoriteit Persoonsgegevens (n 50) 12.

¹⁰⁸ European Commission, ‘Proposal for a Regulation of the European Parliament and the Council amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus)’ COM(2025) 837 final.

¹⁰⁹ Digital Omnibus (n 108) 1.

¹¹⁰ Digital Omnibus (n 108) 6.

¹¹¹ Proposal for a Digital Omnibus Regulation, art 88 (c).

¹¹² Proposal for a Digital Omnibus Regulation art 3 (3).

¹¹³ See Noyb, ‘Digital Omnibus: The EU Commission wants to wreck core GDPR principles’ (*Noyb.eu*, 19 November 2025) <https://noyb.eu/en/digital-omnibus-eu-commission-wants-wreck-core-gdpr-principles> accessed 7 January 2026.

¹¹⁴ ‘EDPB-EDPS Joint Opinion 2/2026 On the Proposal for a Regulation as regards the simplification of the digital legislative framework (Digital Omnibus) (EDPB, 10 February 2026) 17.

information for the data subjects about the processing activities of controllers.¹¹⁵ Besides informing the interpretation and application of the regulation as a whole, it translates into different kinds of obligations for controllers.¹¹⁶ These provisions differ according to the intrinsic or instrumental value of transparency they advance, the target beneficiary of enhanced transparency (individual data subjects, enforcement authorities, society at large), the temporal dimension of processing (ex-ante vs ex-post transparency), and their aims (protective vs empowering).¹¹⁷ Our focus lies on those obligations promoting ex-ante transparency, namely the information duties controllers have to fulfil under arts 13 and 14, in conjunction with art 12 GDPR. These articles mandate certain information duties for controllers towards data subjects prior to the initiation of the processing activities. Their legal and ethical rationale is to raise data subjects' awareness as to qualitative and quantitative aspects of processing in order to enable the effective exercise of their data protection rights.¹¹⁸

Art 12 GDPR dictates the conditions for compliance with substantive information duties, whether ex-ante or ex-post. Accordingly, the relevant information should be provided in writing, including digitally, in a 'concise, transparent, intelligible and easily accessible form, using clear and plain language'.¹¹⁹ This should occur free of charge.¹²⁰ So long these conditions are fulfilled, controllers are free to determine the appropriate measures. This choice usually falls with the publication of privacy policies, cookie-consent forms, information sheets, and terms and conditions on the controllers' website.¹²¹

Arts 13 and 14 GDPR, instead, substantively define the information categories controllers should communicate to data subjects. The former addresses controllers who have directly collected personal data from the data subjects, while the latter only refers to information duties of controllers who have accessed personal data through a third-party. Both articles contain a long list of required disclosures, with art 14 GDPR additionally encompassing the categories of personal data processed and the source of collection.¹²² An important difference between the two provisions is that art 14 contains a broader list of possible derogations from its disclosure obligations.¹²³

3.4.1 Transparency requirements applied to tech-publisher partnerships

The question remaining is then twofold: what disclosures should take place about tech-publisher partnerships and by what means? We build again on the distinction between training and RAG.

As direct collectors of personal data contained in journalistic content, publishers are subject to the obligations under art 13 GDPR, whereas when acting as indirect recipients of personal data from third-party sources, they fall under the obligations of art 14 GDPR. Disclosure should be required with respect the processing activities over which they determine purposes and means, as also confirmed by the CJEU.¹²⁴

The content of the disclosure should include, among other things, the purpose of processing and legal basis, as well as the relevant legitimate interest if this is relied on as the legal basis. In case of training and fine-tuning, publishers would only be required to disclose this information, prior to the moment of transfer, with respect to the transferring of the data, not for the training process itself. That would instead be the responsibility of the LLM-provider under art 14 GDPR, as an indirect collector of personal data.

¹¹⁵ GDPR, recital 39.

¹¹⁶ Jef Ausloos and Michael Veale, 'Researching with Data Rights' (2020) *Technology and Regulation*, 141.

¹¹⁷ Jef Ausloos and Michael Veale (n 116) 141-144.

¹¹⁸ Radim Polčák, 'Article 12. Transparent information, communication and modalities for the exercise of the rights of the data subjects' in Christopher Kuner and others (eds) *The EU General Data Protection Regulation (GDPR)* (Oxford University Press 2020), 402.

¹¹⁹ GDPR, art 12 (1).

¹²⁰ GDPR, art 12 (5).

¹²¹ Arianna Rossi and Gabriele Lenzini, 'Transparency by-design in data-informed research: a collection of information design-patterns' (2020) 37 *Computer Law and Security Review*, 4.

¹²² See GDPR, art 13-14.

¹²³ GDPR, art 14 (4). Controllers falling under art 13 GDPR would only be exempted from disclosure obligations if they can prove that the data subject already has the relevant information, GDPR, art 13 (4). Art 14 (5) (b) for instance also grants a derogation for controllers where compliance would render impossible or seriously impair the achievement of the processing objectives.

¹²⁴ See *Fashion ID*, para 106.

When it comes to RAG, instead, publishers remain jointly in control of means and purposes of OpenAI's processing activities. This means that, in complying with arts 13 and 14 GDPR, publishers should be able to account for the processing necessary for the purposes defined in the joint controllership agreement prior to making the article available for retrieval. In this scenario, LLM-providers could rely on art 14 (5) (a) GDPR exempting indirect controllers from disclosure obligations where data subjects have already received sufficiently detailed information (i.e. from the direct controller). This should not count for additional uses of that data that LLM-providers could make outside the scope of the partnership, which they should duly explain.

Considering the remote feasibility of individually approaching all data subjects affected by the licensing of personal data to LLM-providers, we see three cumulative steps through which compliance could be achieved. This aligns with the suggestion advanced by both the EDPB and CJEU of pursuing a multi-layered approach to fulfil transparency duties.¹²⁵

First, through a general privacy policy. Here, it is interesting to note that the Italian DPA complained to GEDI about the inadequacy of solely a privacy policy to reach the subjects of reporting, considering this is typically addressed to website visitors.¹²⁶

Secondly, when collecting personal data on the basis of consent (i.e. journalistic interviews), journalists could explicitly mention the likelihood of subsequent processing by an LLM-provider. This would not cover non-consensually obtained personal data in the normal course of journalistic activity and, in case of training and fine-tuning, it would still require LLM-providers to fulfil their duties under art 14 GDPR. This could be achieved through the third step we suggest, namely public awareness campaigns. These are not mentioned in art 13 GDPR but they are under art 14 (5) (b) GDPR as an appropriate safeguard controllers could take where the provision of that information would prove impossible. While this should be seen as an exception which controllers should narrowly interpret,¹²⁷ relying on public campaigns as an additional measure might prove more effective than privacy policies. This is particularly relevant for raising awareness among (potential) data subjects about the new implications of having one's personal data collected through journalistic reporting, given the emerging institutional ties between publishers and LLM-providers. Article 29 Working Party suggests these campaigns could rely on online and print advertisements.¹²⁸ The press releases announcing the partnerships could have been a first step to this end, instead of serving PR and marketing goals. Although one can evince from these announcements that some form of data processing involving publishers' journalistic content will take place, these lack the clarity and granularity that would be required for proper disclosures in light of art 12-14 GDPR.

3.5 The journalistic processing regime under art 85 (2) GDPR

Being journalistic data the object of these partnerships, we should enquire at last to what extent these may benefit from the special regime under art 85 (2) GDPR protecting journalistic expression.

The provision obliges EU Member States to exempt individuals and organisations engaged in journalistic activities from several data protection obligations. These can include most of the obligations in the GDPR, including the applicability of data protection principles, data subject rights, controller and processor duties, and international transfer of personal data.¹²⁹ As stated in the provision, the exemption is meant to reconcile the right to personal data protection with that of freedom of expression and information. In practice, this delegation to national legislators has resulted in significant legal heterogeneity across the Member States.¹³⁰ For instance, the substantive GDPR obligations concerning lawfulness and transparency of processing

¹²⁵ EDPB (2024), 29.; Case C-422/24 *Integritetsskyddsmyndigheten v AB Storstockholms Lokaltrafik* [2025] ECLI:EU:C:2025:980, para 42.

¹²⁶ Garante per la Protezione Dati (n 8).

¹²⁷ 'Guidelines on transparency under Regulation 2016/679' (Article 29 Working Party, 11 April 2018) 31

¹²⁸ Article 29 Working Party (n 127)

¹²⁹ GDPR, art 85 (2).

¹³⁰ See David Erdos, *European Data Protection Regulation, Journalism, and Traditional Publishers: Balancing on a Tightrope?* (Oxford University Press 2019); Kamrul Faisal, 'Journalism vs. Data privacy: The GDPR dilemma in reporting crimes' (*Internet Policy Review*, 25 July 2024) <https://policyreview.info/articles/news/journalism-vs-data-privacy-gdpr-dilemma/1787> accessed 7 January 2026.

discussed above, are dealt with quite differently under the journalistic exemption implemented at national level. Some countries, like the Netherlands, exempt journalistic activities from most of the GDPR, including transparency obligations, yet not from the need for a legal basis.¹³¹ Some German Länder go beyond that, excluding even the application of the lawfulness obligations, whereas the Italian Privacy Code does not seem to exempt journalistic processing from either of them.¹³²

3.5.1 Defining journalism under the GDPR

The qualification of ‘journalistic’ purpose is instead a matter of interpretation by the CJEU as a uniform concept of EU law. Trends in the case law hint to an expansive direction, both in terms of the personal scope of the exemption (who qualifies as a journalist?) as well as of the material scope (what is a journalistic activity?).¹³³ As such, for the CJEU, journalistic activities ‘[...] have as their purpose the disclosure to the public of information, opinions or ideas, irrespective of the medium which is used to transmit them’.¹³⁴

Hence, because the CJEU, in line with ECtHR,¹³⁵ deliberately avoids restricting the exercise of journalistic freedom to a particular medium, there is no reason to exclude the journalistic use of LLMs and LLM-based applications from the scope of this protection, the moment these are used as media technology.¹³⁶ However, when considering the application of Art 85 (2) to licensing deals for journalistic content, there are still two issues requiring further scrutiny. Firstly, the fact that the processing activities at stake do not contribute to the production of journalistic content, only to its dissemination; secondly, the journalistic relevance of their purpose.

The journalistic activity

The licensing of journalistic content carried-out under tech-publisher partnerships is not necessary for the production and publication of news; rather, it may only serve its further dissemination. According to Sarah Eskens, applying art 85 (2) GDPR to news dissemination would be inconsistent with the telos of the provision, namely preventing a chilling effect on the exercise of journalistic freedoms to write and publish news.¹³⁷ However, although the GDPR does not expressly state this, the CJEU did not exclude the possibility that the special regime could apply to the dissemination of information extracted from publicly available tax records via mobile text-based services, provided that the underlying purpose would be journalistic.¹³⁸ Consistently with the human rights protection afforded to journalists for news dissemination and the press’ ‘valuable secondary role in maintaining and making available to the public’ news archives, we should not dismiss the application 85 (2) GDPR to processing activities aimed at news dissemination.¹³⁹

¹³¹ Uitvoeringswet Algemene Verordening Gegevensbescherming, art 43 (2).

¹³² Bayerisches Datenschutzgesetz (BayDSG) vom 15. Mai 2018 (GVBl. S. 230, BayRS 204-1-1), art 38 (1); Gesetz zum Schutz personenbezogener Daten im Land Brandenburg (Brandenburgisches Datenschutzgesetz - BbgDSG) vom 8. Mai 2018 (GVBl. I/18, [Nr. 7]), § 29(1); Decreto Legislativo 10 agosto 2018, n. 101, recante “Disposizioni per l’adeguamento della normativa nazionale alle disposizioni del regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE” (in G.U. 4 settembre 2018 n.205), art 137.

¹³³ Cecile de Terwangne and Alejandra Michel, ‘Processing of personal data for “journalistic purposes” in *Deep diving into data protection: 1979-2019: celebrating 40 years of research on privacy data protection at the CRIDS* (Larcier 2021) 208.

¹³⁴ The CJEU also elaborated a detailed list of criteria to consider on a case-by-case basis, including how the data processing contributes to a debate of public interest, the publicness and prior conduct of the data subject(s) affected, the subject of the reporting, the form, consequences and veracity of publication. Case C-345/17 *Buvids* [2019] ECLI:EU:C:2019:122, para 63-66.

¹³⁵ See *Times Newspapers Ltd v The United Kingdom* App no 3002/03 and 23676/03 (ECtHR, 10 March 2009) para 27.

¹³⁶ See Gionata Bouché and others, ‘LLMs as media technology: implications under article 10 ECHR’ (2026) 33 *International Journal of Information Law and Technology* 1; The CJEU also acknowledged that publishing a video on social media platforms, such as YouTube, qualifies as an activity encompassed within the scope of the journalistic exemption. *Buvids*, para 69.

¹³⁷ Sarah J Eskens, ‘A right to reset your user profile and more: GDPR-rights for personalized news consumers’ (2019) 9 (3) *International Data Privacy Law* 153, 159-160.

¹³⁸ Case C-73/07 *Tietosuojavaluuttetu v Satakunnan Markkinapörssi Oy and Satamedia Oy* [2008], ECLI:EU:C:2008:727, para 62.

¹³⁹ See *Jersild v Denmark* App no 15890/89 (ECtHR, 23 September 1994) para 35; *Times Newspapers Ltd v United Kingdom* App no 3002/03 and 23676/03 (ECtHR, 10 March 2009) para 45.

The journalistic purpose

The key question remains whether the purpose underlying activities in tech-publisher partnerships is journalistic. Art 85 (2) GDPR can only apply when the relevant processing activity ‘solely’ fulfils a journalistic purpose.¹⁴⁰ This already appears from art 9 Data Protection Directive, pre-dating art 85 GDPR.¹⁴¹ The drafters of the GDPR omitted this reference from art 85 (2) GDPR, though they did include it in the corresponding (non-binding) recital.¹⁴² In its case law examining art 9 DPD, the CJEU does explicitly write ‘solely’ when discussing the applicability of the provision.¹⁴³ However, the CJEU does not understand this as implying exclusivity – parallel and collateral objectives may be sought through the processing activity, including the pursuit of commercial profit.¹⁴⁴ No judgments on the subject matter are yet available post-GDPR’s entry into force, but it is unlikely that the CJEU will revert from its interpretation of art 85 (2) GDPR.¹⁴⁵

Still, this does not mean that any news-disseminating activity by commercial entities would meet the relevant threshold. For instance, despite contributing to the dissemination of information to a wide public through its search engine, Google could not rely on the journalistic exemption.¹⁴⁶ OpenAI, a prolific dealmaker, and its media partners have consistently stressed the contribution of these collaborations to the promotion of ‘high-quality journalism’.¹⁴⁷ Indeed, these deals may open new frontiers for the reach and distribution of journalistic content. Mathias Dörpfner, CEO of media conglomerate Axel Springer, made this explicit in the press release announcing the companies’ deal:

‘[...]We want to explore the opportunities of AI empowered journalism – to bring quality, societal relevance and the business model of journalism to the next level.’¹⁴⁸

Similarly, the Italian GEDI states that licensing its journalistic content is intended to support ‘developing new and innovative ways to make editorial content accessible through AI-based tools’.¹⁴⁹ Of course, we cannot rely solely on these companies’ PR-lingo understating the relevant commercial interests, notably product improvement for LLM providers and financial compensation for publishers. At the same time, these deals do provide in fact an innovative framework for news dissemination. This means that we should not pre-emptively dismiss the application of art 85 (2) GDPR; instead, we should tackle these new normative tensions concerning the qualification of ‘journalistic processing’ on a case-by-case basis.

In more general terms, we see an implicit link between the existence of a journalistic purpose and the type of processing operation conducted. Deals involving the licensing of journalistic content for pre-training or fine-tuning of LLMs seem to pursue a too generic end. Improvement of the model in the training phase could virtually support any use-case, across all sectors. While journalistic content might enrich the models with accurate information on current past events, that is only collateral. The information merely serves as

¹⁴⁰ Herke Kranenborg, ‘Article 85 Processing and freedom of expression and information’ in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020).

¹⁴¹ Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281/31, art 9.

¹⁴² GDPR, recital 153.

¹⁴³ *Satamedia*, para 50-59; See *Buvids* para 56-62.

¹⁴⁴ *Satamedia*, para 59; Kranenborg (n 140).

¹⁴⁵ In his opinion to a (at the time of writing) pending request for preliminary ruling, AG Szpunar confirms this stance. Opinion of Advocate General Szpunar, Case C-199/24 *ND v Legal Newsdesk Sweden AB* [2025] ECLI:EU:C:2025:670, para 43.

¹⁴⁶ Case C-131/12 *Google Spain*, para 85.

¹⁴⁷ See for instance GNM press office, ‘Guardian Media Group announces strategic partnership with OpenAI’ *The Guardian* (14 February 2025) <https://www.theguardian.com/gnm-press-office/2025/feb/14/guardian-media-group-announces-strategic-partnership-with-openai>; Mistral AI team, ‘Purr-fectly informed’ (*Mistral*, 16 January 2025) <https://mistral.ai/news/mistral-afp> accessed 7 January 2026.

¹⁴⁸ See AxelSpringer, ‘Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism’ (axelspringer, 12 December 2023) <https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism> accessed 7 January 2026; Louis Dreyfus and Jérôme Fenoglio, ‘Le Monde and OpenAI sign partnership agreement on artificial intelligence’ *Le Monde* (16 May 2025) https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html accessed 7 January 2026.

¹⁴⁹ GEDI Group, ‘Press Release Data Protection Authority’ (*GEDI*, 29 November 2024) <https://www.gedi.it/en/media/news/gedi-group-press-release-data-protection-authority> accessed 7 January 2026.

high-quality data from a natural language processing point of view. The societal function of news content matters little in this context. The same, though possibly to a lesser extent, goes for fine-tuning, given that improved dissemination of news could be one of the myriads of objectives that could be achieved through these processing activities.¹⁵⁰ In both cases, the commercial objectives pursued seem to overshadow the journalistic aims so substantially that the latter are relegated to an ancillary role.¹⁵¹

This differs, however, when news content is retrieved through RAG, integrating journalistic content into the output generation process without additional training. The processing does not reduce news articles to raw data points; rather these are analysed to capture their societal function. The journalistic purpose of news dissemination, here, seems more central than in the case of pre-training and fine-tuning. We are inclined to consider this processing arrangement as potentially falling within the scope of the journalistic exemption. This presumption should be strengthened where the decision to enter these partnerships is made by, or at least is substantially influenced, by editorial departments, and not just by commercial managers. This cooperation would differ, in our opinion, from third-party indexing carried out by search engines like Google independently from the activities of publishers and its editorial teams, previously excluded by the CJEU from the special processing regime.¹⁵²

It should be emphasised, again, that relying on art 85 (2) GDPR does not automatically entail the release of controllers and processors involved in the partnerships from all data protection obligations listed in the provision. National legislators decide which exemptions to grant. This special regime was never conceived as an obligation on Member States to create absolute derogations for journalistic processing;¹⁵³ rather, national implementations should provide these only when strictly necessary, ensuring balance between media freedom and personal data protection in light of the EU Charter and the CJEU's case law.¹⁵⁴

The exemption from lawfulness and transparency principles remains therefore highly contingent on national law. In any event, it could not apply to model training, but only to processing operations like RAG that strictly reflect the aim of journalistic dissemination under the GDPR.

4. Conclusion

The number of tech-publisher partnerships is on the rise. LLM-providers are in fact changing the online information retrieval landscape and possibly the future of online news. Concrete opportunities for financial gain, promises of innovation in news dissemination, mixed with a good dose of FOMO, are luring especially large publishers into these deals.

In this article, we tried to show how the flamboyant press releases announcing these ventures mask complex GDPR-related questions with excessive simplifications. The GEDI-Open AI case is a textbook example of this. With no detail of the relevant processing operations, partnerships assumed anything could pass as the pursuit of journalistic innovation, as long as a news publisher and its content are involved. Instead, our analysis so far reveals the intricacy of the tensions raised by these new understandings, requiring careful evaluations in the compliance process.

Our main findings so far hint to the importance of distinguishing the available technical solutions (pre-training and fine-tuning vs RAG) in order to assess the genuine nature of the alleged processing purposes

¹⁵⁰ Of course, unless fine-tuning would be carried out for the fulfilment of a specific journalistic task (i.e. news summarisation).

¹⁵¹ The Dutch DPA takes a similar view: 'Some generative AI-applications could contribute to the provision of news to the public at large. However, due to the wide deployability of foundation models, it can be argued that the function of these models is not exclusively confined to the distribution of news to the public at large'. *Autoriteit Persoonsgegevens* (n 50), 12.

¹⁵² See *Google Spain* para 85.

¹⁵³ See for instance *Verfassungsgerichtshof* (Austria), *Erkenntnis G 287/2022*, *G 288/2022* vom 14. Dezember 2022 (14 December 2022), paras 52-54.

¹⁵⁴ David Erdos, 'Special, Personal and Broad Expression: Exploring Freedom of Expression Norms under the General Data Protection Regulation' (2021) 40 *Yearbook of European Law* 398, 408; The CJEU also stressed this point in *C-73/07 Satamedia* para 56.

and legitimate interests pursued. It is already clear that if RAG can offer an equally effective solution for the purpose of news dissemination through LLM-based systems, it will more easily satisfy the requirements of the GDPR. Moreover, while in our view this kind of processing could potentially benefit from the exemptions under art 85 (2) GDPR, national legislators and data protection authorities would need to further elaborate on the required balance in light of the EU CFR. As this cannot lead to a complete disapplication of the GDPR under national law, we invite the relevant authorities, including national legislators and DPAs, as well as the contractual parties in their joint controllership agreement, to carefully consider how to strike the relevant balance in relation to this new frontier of news dissemination.



Copyright (c) 2026, Gionata Bouché, Marten Steketee.

Creative Commons License. This work is licensed under a Creative Commons Attribution-Non-Commercial-NoDerivatives 4.0 International License.