Legal considerations in the use of synthetic data for AI development and finetuning: The case of LLMs4EU

Kossay Talmoudi, Khalid Choukri, Amélie Gourgeot, Florine Astruc
ELDA, ALT-EDIC
9 Rue des Cordelières 75013 Paris, 1 Pl. Aristide Briand 02600 Villers-Cotterêts
{Kossay, Khalid, Amelie}@elda.org, Florine.astruc@alt-edic.eu

Abstract

This paper examines the legal implications of using synthetic data to develop and fine-tune general-purpose AI models in the European Union, using the LLMs4EU project as a case study. It situates synthetic data within the Union's broader data policy and highlights it as a candidate tool for reconciling data availability with regulatory constraints. From a data protection perspective, it analyses whether and when synthetic data should be classified as "personal data" under the GDPR. From a copyright and contractual standpoint, the paper assesses the risks that synthetic datasets may embed infringing content or derive from other models, in light of the GEMA v. OpenAI ruling on memorised works and emerging analyses of liability for AI-generated outputs, and considers the constraints imposed by model licensing and acceptable-use policies on using models to generate training data for other models. The paper concludes that synthetic data can play a valuable role in mitigating legal risks and enabling compliant AI development in LLMs4EU, but only if its generation and use are embedded in robust governance frameworks that address data protection, copyright and contractual obligations across the entire data value chain.

Keywords: Synthetic data, finetuning, training data

1. Introduction

LLMs4EU is an EU-funded project that aims to fine-tune general-purpose AI models capable of addressing concrete, domain-specific cases, with a particular consideration given to linguistic diversity. This objective presupposes access to diverse datasets, which in turn raises recurrent legal and practical challenges concerning both personal and non-personal data. In this context, the project considers synthetic data as an important component of the training mix, complementing human-generated data and potentially reducing the dependency on scarce or legally constrained datasets.

Recent technical work has stressed that, under current trajectories, the stock of publicly available human-generated text will be insufficient to sustain large-scale LLM training, with projections of exhaustion between now and 2032 if present trends continue (Villalobos et al., 2024). This scarcity is a focal point for the Union's data policy, in which the 2020 European Strategy for Data and the subsequent Data Union Strategy of 2025 urge for a systematic increase in data availability for innovation and competition. Such efforts to limit data scarcity are also reflected in the adoption of data space initiatives such as the Language Data Space.

In the framework of the legal bundle referred to as the EU data laws, the Data Act adopts a broad and technology-neutral understanding of "data" in its article 2(1), agnostic of form, source and structure, and encompasses personal and non-personal data, thereby opening the door to the application of multiple legal regimes depending on provenance and use. Synthetic data are not therefore specifically defined in this legal instrument, but such agnostic definition englobes it.

Synthetic data is increasingly presented as a means to reconcile the objective of quality data and the respect of legal constraints. In LLMs4EU, synthetic data thus appears as a technical response to data scarcity and as a potential regulatory tool to mitigate legal risks mainly related to data protection and intellectual property rights.

The turn to synthetic data also reflects the sensitivity around copyright in the context of generative AI training although emphasis on synthetic data as a possible way to reduce reliance on protected material is confronted to the fact that such data presents its set of legal challenges. LLMs4EU must therefore navigate this regulatory landscape and identify under which conditions synthetic data can lawfully support model development in the Union.

2. Synthetic data as an opportunity for AI developers

2.1 Definitions of synthetic data

Synthetic data can be defined in functional terms as data generated by an algorithm that statistically resembles real-world datasets but does not directly reproduce any specific record. It serves multiple functions such as filling gaps where data is missing or cannot be accessed, or to exhibit particular distributions or properties that are hard to obtain in practice. Synthetic data has been for example used operationally for debiasing through techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic examples of under-represented classes in supervised learning.

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 86–90

Contemporary synthetic data for AI training is predominantly generated using deep-learning models. In the context of LLMs, synthetic text can be used to augment existing corpora, to create specialised or domain-specific datasets, and to explore hypothetical scenarios that would be difficult to document otherwise.

Recently, so-called data distillation approaches have been proposed to compress large training corpora into smaller, high-utility synthetic sets. In such approaches, models are trained or fine-tuned on structured synthetic datasets generated by a base model with the aim of preserving performance while reducing dataset size. Such synthetic data can be generated in an environment in which a large language model is prompted to produce outputs that are then re-used as training inputs, either for itself (self-training) or for other models. Such iterative loops promise cost-effective scaling of training data. It is to be noted that empirical work in the LLM field suggests that carefully curated synthetic data can, under certain conditions, support further training or fine-tuning of models, although repeated training on model-generated data can also introduce biases and degradation (Shumailov et al., 2024)

2.2 Synthetic data as a mitigator of legal risk in AI training

2.2.1 Synthetic data in support of data protection compliance

Advanced analytical techniques have shown that even heavily processed and cleaned datasets used to train AI may contain sufficient information to re-identify individuals or to infer sensitive attributes, particularly when combined with other data. This undermines the effectiveness of traditional anonymisation methods and supports the view of supervisory authorities and the European Data Protection Board (EDPB) that anonymisation must be robust and context-specific to prevent re-identification, since in cases where there is a mere prevention of attribution, the data cannot be considered as anonymous (European Data Protection Board, 2025). The question therefore remains whether synthetic data constitute “personal data” within the meaning of Article 4(1) General Data Protection Regulation (GDPR) where they allow information relating to an identifiable person to be inferred. Synthetic data is often proposed as a potential solution, as it can, in principle, either be fully anonymous in fully synthetic data, or at least cut the link between records and identifiable persons while preserving utility in the case of partially synthetic data (Zhang et al., 2022).

Some researchers (Gal and Lynskey, 2023) have argued that synthetic data challenges the conceptual foundations of data protection law because, even where a synthetic dataset no longer contains any record corresponding to an actual individual, it can still be used to make decisions and inferences that affect real persons. Their analysis suggests that many synthetic datasets should be treated as personal data whenever they allow for individual-level impacts, even if the data points themselves are artificially generated.

In all cases, synthetic data produced via other models are dependant in its compliance on whether the underlying model is trained lawfully or not. The EDPB’s opinion on AI models emphasises that developers must ensure that models are not trained on unlawfully processed personal data and that data protection principles apply across the lifecycle of AI models, including training, validation and deployment (European Data Protection Board, 2024). The Joint Research Centre’s work on synthetic data in digital finance indicates that, with appropriate rules for generation, synthetic datasets can yield analytical results closely aligned with those obtained from original data while ensuring compliance with confidentiality requirements (European Commission’s Joint Research Centre, 2024).

Synthetic data, depending on how its generation is carried out, must not be regarded as a way to circumvent data protection constraints, but rather as a privacy-enhancing instrument that aims to minimise privacy risks, provided that its use remains grounded in a compliant governance framework, as clear risks remain to be mitigated, as analysed in section 3.3.

2.2.2 Mitigating copyright risks through synthetic data

Synthetic data may mitigate copyright risk in LLM fine-tuning because it can reduce direct dependence on protected material. Where a fine-tuning corpus is built from human-authored works, the legal exposure arises principally from the possibility that the training set contains protected expression, whether in full works, substantial parts, or fragments that remain recognisable in downstream outputs. This is sensitive in generative AI, where the model may reproduce or closely imitate protected sequences, and where the line between lawful learning and infringing reproduction is often contested (Tyagi, 2025).

Synthetic data offers a partial answer to the copyright question because, when it is generated from a model rather than from direct reuse of protected expression, it is less likely to replicate copyrighted works verbatim or to preserve expressive choices that are protected by copyright. In other words, the legal risk is not eliminated, but it is shifted: the compliance question moves from the downstream corpus to the upstream generation process. If the synthetic corpus is generated by model, thus without directly copying protected text, and if it is sufficiently transformed so that it no longer

contains substantial parts of any pre-existing work, it is generally less problematic from a copyright perspective than a corpus assembled by scraping or reusing protected works directly.

That said, synthetic data is not a categorical panacea. If the synthetic generator itself is trained on copyrighted material without legal scrutiny, or if it emits outputs that reproduce protected expression with sufficient similarity, the resulting corpus may still carry copyright risk. This is why LLMs4EU should treat synthetic data as a risk-reduction technique rather than as a substitute for copyright clearance. The relevant operational question whether the outputs are sufficiently detached from protected source works is analysed in section 3.3.

3. Legal hurdles to the adoption of synthetic data in AI development

Synthetic data used in LLM training lies at the intersection of several legal regimes that may apply cumulatively. From a data protection perspective, the qualification of synthetic datasets as personal or non-personal data determines the applicability of the GDPR. From a copyright perspective, synthetic data derived from models trained on protected works may constitute adaptations or reproductions, and their generation may require appropriate legal scrutiny. In addition, contractual provisions in model licences and data-sharing agreements can impose further restrictions on the generation and reuse of synthetic data, especially in relation to acceptable uses and redistribution.

3.1 Model licensing and acceptable uses

Many foundation models are distributed under licences that include acceptable use policies prohibiting certain applications, such as generating abusive content, violating privacy or using the model to develop competing models. Such provisions may explicitly or implicitly restrict the generation of large synthetic corpora for downstream model training or may condition such uses on obtaining additional permissions. For LLMs4EU, which contemplates using existing models to generate synthetic data, careful scrutiny of such acceptable use clauses is necessary to ensure that the planned uses do not constitute misuse and thus do not result in a breach of contractual obligations.

The emergence of “open-source” or “open-weight” models has raised debates about what openness entails in the AI context, including whether there are restrictions on commercial use, re-distribution of weights, or model-as-a-service deployment. Even where model weights are openly accessible, associated licences may limit training on model outputs or prohibit using the model to generate data that is then used to train another model that competes with the original. In LLMs4EU, relying on such models to create synthetic training data requires an analysis of licence compatibility, particularly when the aim is to finetune AI models that may themselves be released under open policies.

In addition to model licences, and depending on the mechanism of data synthesis, licences governing the data used to generate synthetic data may impose downstream obligations. Where synthetic data is generated from licensed datasets, the question arises whether the synthetic corpus is a derivative work or whether it falls outside the scope of the licence. Given that some rights-holders and collecting societies take the view that output generation constitute reproductions and adaptations, contractual terms may attempt to extend protection to synthetic derivatives as well. In LLMs4EU, both the licensing of input data and the allocation of rights and responsibilities over synthetic outputs must therefore be considered.

3.2 Synthetic data constituted of copyright infringing outputs

AI-generated outputs that closely resemble pre-existing works used in training can give rise to copyright infringement, and liability may attach both to the user who inputs the prompts and to the developer or provider who made the model available (Rosati, 2025). For LLMs4EU, this entails that synthetic data used for further training must be assessed in terms of the concrete risk that it contains infringing sequences that could be reproduced or amplified in downstream models. This entails that the models used to generate the synthetic data need to be assessed on whether they are resilient enough to possibilities of them “leaking” the original data it was trained on, which was feasible with early generative AI models that were not subjected to sufficient alignment procedures (Carlini et al., 2021).

The GEMA v. OpenAI decision is an important case in this regard as it clarified that AI models can embody protected works in a way that triggers copyright liability. The Munich Regional Court held that specific song lyrics were “physically fixed” in the model, that they could be indirectly perceived through prompts, and that this memorisation constituted reproduction within the meaning of Article 2 of the InfoSoc Directive. The court further rejected defences based on quotation, parody or other limitations and ordered OpenAI to provide information and pay damages, finding that the company acted at least negligently despite legal uncertainty.

Synthetic datasets used for training may consist wholly or partly of outputs generated by models that have been trained on unlicensed or infringing

data. In such cases, even if the synthetic data does not contain verbatim reproductions, it may still be tainted by the initial unlawfulness of the training process or may in practice reproduce protected expression when prompted in specific ways. In this sense, research on memorisation and extraction in large language models shows that generative systems can regurgitate training material (Ahmed et al., 2026). The possible persistence of such behaviour makes it necessary for LLMs4EU test synthetic corpora for near-duplication, long-span overlap, and other forms of recoverable expression before relying on them for fine-tuning. Indeed, a synthetic dataset derived from other models cannot automatically be regarded as free from legal constraints.

The legal analysis of synthetic data in LLMs4EU focuses on the specific layer of the value chain at which synthetic outputs are generated and reused. Even when the consortium does not directly process the original human-generated data when creating synthetic datasets, it may still incur responsibility for using synthetic data that embed protected expression or personal information in a way that is functionally equivalent to using the original data.

3.3 Residual data protection risks in synthetic data

When synthetic data is generated by models trained on datasets containing personal data, traces of that data can be found in the output. Synthetic data can therefore be deemed "not inherently anonymous" (Achterberg et al., 2025). This is true as synthetic data can still encode patterns that are traceable to specific individuals.

Privacy evaluations assessed on synthetic data shows that the reidentification risk is residual, at least compared to real world datasets: A study conducted on synthetic data in the field of child and adolescent mental health performed three attack-based privacy evaluations on the synthetic data used. The evaluation shows that while "the overall risk remains quite low", "there is some potential for linking data to individuals" (Haizoune et al., 2026). From a legal point of view, such risk is sufficient for the GDPR to apply. Such research is also in line with previous findings that distinguish between partially synthesised data that remains “vulnerable to membership inference” and fully synthesised data that remains quite resilient when met with adversarial attacks (Zhang et al., 2022).

For LLMs4EU, this implies that synthetic datasets must be subjected to rigorous privacy assessment and testing to demonstrate whether there are risks in reidentifying the legal persons in synthetic datasets. In practice, this assessment should rather include structured tests of whether individuals can still be singled out. A first step is to evaluate the generation process itself: the consortium should verify whether the synthesis method preserves rare combinations, exact values, or patterns that could disclose information about identifiable persons. A second step is to conduct re-identification testing, for instance by attempting membership inference, attribute inference, and reconstruction attacks on the synthetic dataset, in order to measure whether an attacker could recover data about specific individuals. A third step is to document all mitigation measures when these are carried out. In this sense, compliance should be demonstrated through clear methodologies and not only a declaration of the synthesised nature of the data.

4. Conclusion

Synthetic data offers significant opportunities for LLMs4EU in terms of mitigating legal risks, challenging data scarcity and enhancing representativeness, particularly for under-served languages and domains. It can reduce direct reliance on personal data and copyrighted works in training and facilitate wider sharing of data. Nonetheless, synthetic data cannot be assumed to be legally neutral; its generation and use remain embedded in the broader framework of data protection, copyright and contractual law.

Given the persistent risks of re-identification, infringing outputs and contractual non-compliance, LLMs4EU should endorse synthetic data as part of its training strategy only under robust safeguards. These include careful selection of models used for generation, ensuring that acceptable use policies and licences permit synthetic data creation for model training and that their own training processes are sufficiently transparent and reliable not to introduce additional compliance risks; rigorous privacy and copyright audits of synthetic corpora; and documentation of generation processes as part of accountability under the GDPR and the AI Act. In particular, using models to generate data that will then train other models should be explicitly permitted and appropriately governed. It needs to be highlighted that when synthetic datasets are treated as legally “easier” to use and share, there is a risk that they will be used in contexts where their limitations are not sufficiently understood, which could lead to potential liability.

The legal status of synthetic data will remain dynamic as courts and regulators confront new cases, including further decisions on AI training and output liability and evolving interpretations of personal data and copyrightability. In LLMs4EU, synthetic data will be thus treated as a component of a broader legal and technical governance framework that is periodically reassessed in light of new case law, regulatory guidance and technical evidence on the behaviour of models trained on synthetic data.

5. Acknowledgements

This work is supported by the LLMs4EU project, funded by the European Union through the Digital Europe Programme (DIGITAL) under the grant agreement 10119847.

6. Bibliographical references

Achterberg, J., van Dijk, B., Waseem, H.M., Gallos, P., Epiphaniou, G., Maple, C., Haas, M. Spruit, M. (2025). The Data Sharing Paradox of Synthetic Data in Healthcare. *arXiv preprint arXiv:2503.20847*.

Ahmed, A., Cooper, A. F., Koyejo, S., & Liang, P. (2026). Extracting books from production language models. *arXiv preprint arXiv:2601.02671.*

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21),* pages 2633-2650.

European Commission, Joint Research Centre, (2024) Synthetic data in the Data Hub of the Digital Finance Platform, *Publications Office of the European Union, Luxembourg*

European Data Protection Board, (2025) Guidelines 01/2025 on Pseudonymisation

European Data Protection Board, (2024) Opinion 28/2024 on certain data protection aspects of AI models

Gal, M. Lynskey, O. (2023), Synthetic Data: Legal Implications of the Data-Generation Revolution, *109 Iowa Law Review*

Goodfellow, J. Pouget-Abadie, J. Mirza, M. Xu, B. Warde-Farley, D. Ozair, S. Courville, A. Bengio, Y. (2014) Generative adversarial networks, In *Proceedings of the 27th International Conference on Neural Information Processing Systems, Volume 2,* pages 2672-2680

Haizoune, M., Leventhal, B. L., Pant, D., Nytrø, Ø., Koochakpour, K., Koposov, R.A., Øhlckers, L.R., Skokauskas, N. (2026). Balancing Privacy and Utility in Child and Adolescent Mental Health Services Research: Retrospective Cohort Study on Synthetic Data Generation. *JMIR Medical Informatics, 14*, e71819.

Rosati, E. (2025) Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law, *European Journal of Risk Regulation 16(2)* p. 611

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y. (2024) AI models collapse when trained on recursively generated data. *Nature 631*, pages 755–759.

Tyagi, K. (2025) Synthetic Data, Data Protection and Copyright in an era of Generative AI, *16 JIPITEC* p.176

Villalobos, P. Ho, A. Sevilla, J. Besiroglu, T. Heim, L. Hobbhahn, M. (2024) Will we run out of data? Limits of LLM scaling based on human-generated data, *arXiv preprint arXiv:2211.04325*

Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. Journal of biomedical informatics, *125*, 103977.