sciendo

Regulating LLM training under EU copyright law and its linguistic implications for smaller languages such as Estonian

Johanna Priske
Pawan Kumar Dutt

Department of Law
Tallinn University of Technology
Ehitajate tee 5
Tallinn 19086, Estonia
jopris@taltech.ee
pawan.dutt@taltech.ee

Abstract: The rapid advancement of artificial intelligence (AI) and text and data mining (TDM) technologies has become a crucial factor in supporting linguistic diversity of smaller languages, such as Estonian. However, the growing reliance on large-scale data processing raises significant legal and ethical challenges, particularly concerning the lawful use of copyrighted materials to train AI systems. While the European Union has introduced regulatory instruments such as the AI Act and the Copyright Directive to address these concerns, the intersection of AI development, TDM, and intellectual property rights remains insufficiently resolved. This research aims to explore how AI-driven text and data mining can both comply with European copyright law and actively contribute to the preservation and wider use of the Estonian language. The authors investigate this research topic from the perspective of the Estonian language, enmeshed within Estonian cultural norms and parameters. The analysis examines the legal and economic appropriateness of enabling the training of generative AI-based large language models (LLMs) in the Estonian language as a means of promoting the language and assuring its appropriate use in digital environments, thereby contributing to its long-term relevance. Methodologically, the study combines a comparative analysis of EU, US, and Asian legal frameworks with a theoretical evaluation of existing models for balancing innovation and rights protection. The article seeks to provide concrete recommendations

https://doi.org/10.2478/bjes-2026-0029 *TalTech Journal of European Studies*, 16(2(45))
Tallinn University of Technology (ISSN 2674-4619)

for developing AI and data mining practices that enhance linguistic innovation while safeguarding creators' rights, thereby contributing to Estonia's digital and cultural resilience within the EU's evolving AI ecosystem.

Keywords: *copyright, Estonian, large language models, low-resource languages, text and data mining*

1. Introduction

The rapid development of generative artificial intelligence has transformed text and data mining (TDM) from a specialised analytical technique into a foundational infrastructure of contemporary knowledge production. Large language models rely on extensive textual corpora to identify statistical regularities in language, enabling them to generate fluent and context-sensitive outputs (Wang *et al.*, 2025, pp. 1956–1958). As a result, TDM has moved to the centre of debates on copyright law, innovation policy, and cultural governance, particularly in jurisdictions that maintain strong traditions of authorial protection (Vardanyan *et al.*, 2026, pp. 2–4).

Within European copyright law, TDM has historically been tolerated as a non-expressive use of protected works, justified by its analytical character and minimal interference with the normal exploitation of those works. This understanding informed the adoption of the TDM exceptions under the Digital Single Market Directive ('Council directive (EU) 2019/790', 2019), which sought to facilitate data-driven research while preserving rightsholders' autonomy through opt-out mechanisms. However, the emergence of generative AI challenges the conceptual assumptions underlying these provisions, as model training entails the large-scale ingestion and internalisation of expressive material rather than transient analysis alone (Quintais, 2025).

It has been argued that generative model training differs qualitatively from classical TDM because it results in durable internal representations that may substitute for protected works in downstream markets (Dornis, 2025).

The regulatory landscape has become further complicated by the adoption of the Artificial Intelligence Act ('Council regulation (EU) 2024/1689', 2024), which introduces a horizontal framework for governing AI systems based on risk, transparency, and accountability. While the AI Act addresses concerns

relating to fundamental rights and market power, it deliberately refrains from harmonising copyright rules governing training data (Söderlund and Larsson, 2024, pp. 9–12). This separation risks producing regulatory asymmetry, whereby AI systems may comply with safety and governance obligations while remaining exposed to unresolved copyright claims at the level of data acquisition.

Beyond doctrinal concerns, the expansion of TDM-driven AI raises broader cultural questions regarding linguistic representation in digital infrastructures. Scholars working on language and AI have noted that data-intensive models tend to privilege languages with extensive digital corpora (Haokip, 2022). In this context, the legal treatment of TDM becomes indirectly linked to questions concerning the long-term visibility of smaller language communities.

Against this background, this article examines the relationship between TDM, generative AI, and copyright law from a comparative legal perspective, while also considering the linguistic and cultural implications of AI training on smaller regional languages. The analysis addresses three central research questions. First, what is the current legal framework governing text and data mining for AI training under European Union copyright law, particularly in light of the Digital Single Market Directive and emerging judicial interpretation? Second, how does the EU regulatory approach compare with legal frameworks in other jurisdictions, such as the United States and selected Asian countries, in addressing the use of copyrighted works for AI training? Third, how can existing regulatory frameworks accommodate the technological realities of AI training without undermining copyright protection or deepening cultural and linguistic inequalities, particularly for underrepresented languages such as Estonian? Methodologically, the analysis combines a doctrinal examination of EU law with comparative insights from the United States and selected non-European jurisdictions, alongside theoretical perspectives from linguistic and communication studies. The aim is to identify legally and culturally sustainable approaches to AI-driven innovation that support both creators' rights and linguistic diversity.

2. The theoretical viewpoint

The rapid development of artificial intelligence systems, particularly LLMs, has intensified existing asymmetries in the digital linguistic landscape. Research on language technologies demonstrates that low-resource languages face structural barriers to AI development, including limited training data and reduced commercial incentives for technological investment (Shompa, 2025). The theoretical perspectives examined in this chapter provide a framework for understanding how technological infrastructures influence language use.

2.1 Ecological approach

The ecological approach to language acquisition was developed most prominently by Leo van Lier (2004). It conceptualises language learning as a process that emerges from the dynamic interaction between learners and their environment rather than as the internal acquisition of fixed linguistic structures.

Van Lier (2004, p. 66) further emphasises the importance of semiotic mediation, arguing that language learning occurs through participation in meaning-making activities involving symbolic systems, linguistic resources, and technological media. Technological environments, therefore, are not neutral channels but active components shaping how language is perceived, used, and acquired. Digital platforms, artificial intelligence systems, and automated communication tools introduce new affordances that influence linguistic behaviour and access to language resources.

Technology expands the range of linguistic affordances available to users by enabling interaction across linguistic boundaries and providing immediate access to linguistic feedback and generated language. At the same time, ecological theory emphasises that these technological affordances are embedded within broader sociocultural systems, which means that their effects on language use and acquisition depend on patterns of access, cultural norms, and institutional structures (van Lier, 2004).

Overall, van Lier's ecological approach provides a theoretical framework for understanding how language development and use are shaped by complex interactions between individuals, technologies, and sociocultural environments.

2.2 Neo-language acquisition and learning theory

Neo-language acquisition theory builds upon classical second-language acquisition models by recognising the transformative impact of digital technologies on how languages are learned, used, and internalised (Eslit, 2023).

A central premise of the neo-language acquisition theory is that digital technologies reshape the cognitive and social processes underlying language acquisition. Interactive technologies such as AI chat systems, language-learning platforms, and automated translation tools provide immediate feedback, adaptive input, and personalised interaction. These features allow learners to engage continuously with linguistic structures and experiment with language. As a result, language learning becomes more iterative, dynamic, and responsive to individual learners' needs, enabling users to develop linguistic competence through repeated interaction with intelligent systems. (Eslit, 2023)

Another key component of the neo-language acquisition theory is the capacity of digital systems to adjust linguistic input dynamically to match learner's current proficiency level. Artificial intelligence systems are capable of generating contextually relevant language, correcting errors, and providing alternative formulations in real time. This creates an environment in which learners receive immediate and context-sensitive feedback, accelerating the language acquisition process and reinforcing linguistic patterns through interaction. (Eslit, 2023)

AI-mediated communication allows learners to consume and produce language simultaneously. Through repeated interaction, learners internalise linguistic patterns, vocabulary, and syntactic structures, while also contributing to the digital linguistic environment itself. (Eslit, 2023)

2.3 Computer-mediated communication theory

Computer-mediated communication (CMC) theories examine how communication technologies shape interpersonal interaction, perception, and linguistic behaviour. Unlike face-to-face communication, CMC relies on technological interfaces that filter or transform communicative cues, influencing how users interpret messages and establish social relationships (Walther, 2011).

A central theory within this framework is the social presence theory, originally developed by Short, Williams, and Christie (1976). Social presence refers to the degree to which users perceive others as real and psychologically present during mediated interaction. The perception of social presence depends largely on the availability of social and non-verbal cues, such as tone, facial expressions, and immediacy of response. Media that provide more cues can foster stronger interpersonal connections, whereas text-based communication typically provides fewer signals and may initially reduce the perceived immediacy of interaction (Wrench, Punyanunt-Carter and Thweatt, 2026).

Even in text-based communication, individuals can perceive strong social presence when they experience meaningful interpersonal interaction, suggesting that social presence is not determined solely by technological features but also by user perception and interaction patterns (Wrench, Punyanunt-Carter and Thweatt, 2026, pp. 386–387).

Closely related to social presence is the media richness theory developed by Daft and Lengel (1986). The theory evaluates communication media based on their capacity to convey information and reduce ambiguity. Media richness refers to a channel's ability to transmit cues and enable feedback. Richer media, such as face-to-face communication, provide greater informational capacity and can facilitate more accurate interpretation of complex or ambiguous messages (Bergin, no date, pp. 2–3).

Media richness theory suggests that the effectiveness of communication depends on matching the richness of the medium to the complexity of the communication task. Rich media are better suited to situations involving ambiguity or emotional nuance, while leaner media may be sufficient for routine or informational exchanges. In digital environments, including AI-based communication systems, text-based interfaces may initially appear less rich. Nevertheless, they offer distinct advantages, including accessibility, scalability, and the capacity to provide structured and adaptive responses. (Bergin, no date)

In the context of AI-mediated communication, these frameworks provide valuable insights into how users perceive AI systems as communicative partners and how technological mediation influences linguistic interaction.

2.4 Linguistic and sociolinguistic perspectives on technology

The emergence of digital communication technologies has prompted extensive debate regarding their impact on language use, structure, and development.

David Crystal (2006) has argued that digital communication, particularly texting, does not undermine linguistic competence. Crystal (2006) argues that digital communication represents a form of linguistic creativity rather than linguistic decline. He emphasises that users of digital media demonstrate high levels of linguistic awareness, adapting language to suit new communicative environments. Digital communication may also strengthen language skills by encouraging users to manipulate linguistic structures, experiment with stylistic variation, and develop new forms of expression appropriate to technologically mediated contexts.

Crystal (2006) further notes that digital language incorporates elements of both spoken and written communication, forming a hybrid linguistic mode. This hybrid form allows users to compensate for the absence of vocal tone and facial expressions by introducing textual strategies that convey emotional and pragmatic information. As a result, digital communication can expand the expressive capacity of written language rather than diminish it.

Building on Crystal's observations, other scholars have examined the emergence of textisms, a term referring to abbreviations, acronyms, and symbolic forms used in digital communication. These linguistic innovations serve functional communicative purposes, allowing users to express emotional tone, conversational nuance, and interpersonal intent despite the limitations of text-based communication. Textisms therefore function as compensatory mechanisms that can enhance communicative clarity (Shortis, 2007).

Both Crystal and Shortis emphasise that digital language follows systematic linguistic patterns and reflects communicative needs rather than linguistic deterioration. The emergence of new linguistic forms in digital environments represents a continuation of historical processes through which language evolves in response to changing communicative conditions.

In the context of AI-mediated communication, these sociolinguistic perspectives provide critical insight into how technological systems influence language use and development. AI systems are trained on digital linguistic corpora that include textisms, abbreviated forms, and hybrid communicative

structures, reinforcing their role as participants in evolving linguistic ecosystems.

2.5 Theoretical expectations and technological realities

Despite the theoretical expectation that technological environments can support linguistic diversity, the development of AI systems reveals a significant distinction between conceptual frameworks and practical implementation. LLM training depends heavily on the availability of extensive digital text, which privileges high-resource languages with a substantial online presence. As a result, languages with smaller speaker populations often remain underrepresented in AI training datasets, limiting the effectiveness of AI tools in those linguistic contexts. Research on extremely low-resource languages demonstrates that while cross-linguistic transfer from high-resource languages may partially improve model performance, such approaches can also distort linguistic representation (Purason, Kuulmets and Fishel, 2025). In practice, the development of language technologies continues to reflect broader structural inequalities in digital infrastructure and resource allocation, leaving smaller languages dependent on targeted policy measures and public initiatives to ensure their continued presence in emerging AI systems. This challenge is particularly relevant for languages such as Estonian, where the development of language technologies relies heavily on coordinated research initiatives and public support to maintain adequate digital representation (Muischnek, 2023).

3. Text and data mining in practice around the world

Having established the theoretical foundations that frame languages as embedded within technological and communicative ecosystems, it becomes necessary to examine how legal frameworks regulate the technological processes that shape these linguistic environments.

3.1 European Union

In the European Union, the regulatory framework for text and data mining was transformed by the adoption of the Digital Single Market (DSM) Directive (Dusollier, 2020). Article 3 concerning the copyright and related rights in the Digital Single Market establishes a mandatory exception

for research organisations and cultural heritage institutions conducting scientific research, while Article 4 introduces a broader exception permitting TDM for any purpose, including commercial uses, provided that users have lawful access and that rightsholders have not reserved their rights ('Council regulation (EU) 2019/790', 2019, Articles 3 and 4).

Both exceptions are conditioned on lawful access, which means that users must legitimately obtain access to the material before engaging in mining activities (Radeisen, 2026, p. 229). The lawful access requirement has been criticised for potentially limiting the practical effectiveness of the exception (Manteghi, 2025).

Article 4 further allows rightsholders to opt out of the exception by reserving their rights in an appropriate manner. However, uncertainty remains as to what constitutes a legally sufficient reservation of rights (Leistner and Antoine, 2025, pp. 1030–1031). Scholarly debate has also arisen over whether large-scale generative AI training properly falls within the concept of TDM (Alonso and Lucchi, 2025), with some authors arguing that such training exceeds the analytical purpose underlying Articles 3 and 4 (Dornis, 2025). These provisions have generated extensive academic debate regarding their practical scope and implementation (Margoni and Kretschmer, 2022).

The regulatory landscape is further complicated by the interaction between the DSM Directive and the EU AI Act (Guadamuz, 2025). This has created tensions and left unresolved questions regarding how copyright compliance in AI training is to be operationalised within the broader EU framework (Quintais, 2025).

One of the earliest cases involving TDM and AI training in the EU was the October 2024 decision of Rechtbank Amsterdam. The District Court of Amsterdam (2024) was called upon to address the legality of automated data processing by a technology operator that used copyrighted materials to train its systems. The Court's reasoning underscored the gap between existing copyright exceptions and the realities of AI training. It emphasised that the statutory TDM carve-outs in the DSM Directive were primarily designed for non-expressive research and scientific purposes, and that the commercial application of generative AI complicates their direct application.

A similar judicial hesitation surfaced in the Regional Court of Hamburg decision of September 2024. While not framed explicitly as a TDM case, the Regional Court of Hamburg (2024) confronted the question of automated reproduction of protected texts in the context of machine training. The Court's

analysis focused on the exclusive rights of authors and the economic value of expressive content, suggesting that large-scale ingestion of such content for commercial exploitation cannot be treated as neutral data processing under the DSM Directive's exceptions.

Another noteworthy development occurred in Hungary, where the Municipal Court of Appeals of Budapest issued a judgement in December 2024, touching on related issues of automated data use. Although the core of the dispute involved digital content distribution rights, the appellate court's reasoning alluded to the implications of large-scale automated processing, emphasising the need to balance innovation with the legitimate interests of rightsholders.

3.2 United States

In the United States, courts address disputes over AI training data under the common-law doctrine of fair use, codified in § 107 of the *US Copyright Act*. Under fair use, courts balance factors such as the purpose and character of the use, the nature of the copyrighted work, the amount used, and the market effect to determine whether use without permission is permissible.

Recent US litigation illustrates how fair use is being applied in the context of AI training. In 'Bartz v. Anthropic' (District Court for the Northern District of California, 2025a), a federal judge in San Francisco found that the use of copyrighted books to train Anthropic's AI model, Claude, could qualify as fair use, likening the training process to a human reader learning from texts. The same decision held that storing millions of pirated copies of books was itself infringement, highlighting that fair use does not automatically legalise all aspects of dataset acquisition.

Similarly, in 'Kadrey *et al.* v. Meta Platforms', a related case alleging unauthorised training of Meta's Llama model on copyrighted books, a US district court dismissed the plaintiffs' claims not because it had established that AI training was necessarily lawful, but because the plaintiffs had failed to develop a sufficiently robust evidential record to counter the defendant's fair use arguments (District Court for the Northern District of California, 2025b).

The reliance on fair use reflects the broader structural flexibility of the US copyright system (Hugenholtz and Senftleben, 2011), which does not explicitly regulate TDM but instead evaluates new technological uses through general copyright principles. This approach allows courts to adapt

existing doctrines to emerging technologies, but it also creates significant legal uncertainty (Chesterman, 2025).

The case-by-case approach does provide flexibility but lacks the predictability that both developers and rightsholders need (Shamov, 2025). While this approach allows courts to accommodate technological innovation, it also means that the legal status of AI training remains unsettled (Thongmeensuk, 2024).

3.3 Asian countries

Unlike the EU's harmonised statutory regime or the US's unified reliance on the fair use doctrine, Asian jurisdictions have adopted diverse approaches reflecting differences in legal traditions, regulatory priorities, and technological policy objectives. The following sections consider selected jurisdictions as illustrative examples of how different legal systems in Asia have approached the regulation of text and data mining and AI training.

3.3.1 Japan

In Asia, Japan presents an example of a jurisdiction that has taken a comparatively permissive approach to the use of copyrighted content for AI training. Under the 2018 revision of Japan's *Copyright Act*, the statute permits data analysis that does not result in the enjoyment of copyrighted expression, a formulation that has been broadly interpreted as encompassing the use of copyrighted works for machine-learning training. In practice, this has been understood to mean that developers can use copyrighted data to train AI systems without explicit authorisation, provided that the use falls within the bounds of lawful data analysis. Additionally, Japan uses the term 'information analysis' rather than 'data mining', giving the concept a broad scope. The data sources subject to TDM are therefore not limited to databases or textual information but may also include linguistic, audio and visual content (Lu and Yuzhi, 2024, pp. 3–4).

By providing greater legal certainty for AI developers regarding the use of copyrighted materials for training purposes, Japan seeks to position itself as a favourable environment for AI research and commercial development (Terui and Chen, 2025, pp. 3–7).

3.3.2 China

China has not adopted specific legislation governing the use of text and data mining technology (Ge, 2025). Instead, courts typically treat TDM activities as a form of web crawling and assess their legality under the general fair use provision contained in Article 24 of the *Copyright Law (China) 2010*, which enumerates twelve statutory exceptions. Where TDM activities satisfy one of these listed exceptions, they may qualify as fair use. Otherwise, they risk being considered infringements. Judicial practice demonstrates that courts may also evaluate large-scale data scraping under unfair competition law, particularly where data constitutes a commercially valuable resource or where technical restrictions against crawler use have been imposed. Because Article 24 of the *Copyright Law (China) 2010* provides a closed list of exceptions, it is often difficult for large-scale TDM activities to fall within its scope, making the legal conditions for lawful data mining comparatively strict and potentially restrictive for technological development (Lu and Yuzhi, 2024, pp. 4–5).

3.3.3 Singapore

Singapore's regulatory model similarly emphasises technological innovation while maintaining legal clarity (Terui and Chen, 2025, p. 8). Singapore has adopted a dedicated statutory framework for computational data analysis (CDA). Sections 243–244 of the *Copyright Act (Singapore) 2021* establish an explicit exception that applies to both commercial and non-commercial uses, provided that the activity qualifies as CDA and the user has lawful access to the material. The Act defines CDA broadly to include the identification, extraction, and analysis of patterns or trends, expressly covering machine learning and related technologies. Additionally, it contains an anti-override rule, under which contractual terms seeking to exclude or restrict permitted CDA uses are rendered void before Singapore courts. However, the exception is subject to important limitations. Its application depends on lawful access, which means that the material must be legitimately acquired, and the exception does not authorise the circumvention of technological protection measures. Separate anti-circumvention provisions prohibit bypassing paywalls, encryption, or other access controls, even where CDA would otherwise be permitted. Moreover, the practical effectiveness of the exception may be weakened by foreign-law and jurisdiction clauses in licensing agreements, creating enforcement uncertainty despite the doctrinal breadth of Singapore's statutory model (Ong, 2025, pp. 7–9).

4. The cultural and social impact of LLMs on smaller regional languages

The deployment of LLMs has significant cultural implications that extend beyond questions of technical performance or economic efficiency. As LLMs are trained on large-scale textual corpora, their linguistic competence reflects existing asymmetries in digital language representation. Languages with extensive online resources are disproportionately well served, while smaller or regional languages remain underrepresented or entirely absent from training datasets. These dynamics risk reinforcing pre-existing hierarchies in global knowledge production, in which dominant languages become further entrenched as the default means of digital communication.

Scholars examining the relationship between artificial intelligence and endangered languages have repeatedly emphasised that data scarcity constitutes one of the primary barriers to meaningful AI support for minority linguistic communities. Haokip (2022) argues that AI linguistic marginalisation by privileging languages that are already institutionally dominant. In this sense, LLMs do not merely mirror linguistic inequalities but may actively deepen them by shaping user interaction patterns, content availability, and the perceived prestige of languages in digital environments.

At the same time, recent research highlights the potential of generative AI to contribute positively to language preservation when models are deliberately adapted to low-resource contexts. LLMs can support documentation, translation, and content generation for underrepresented languages, provided that appropriate datasets and fine-tuning strategies are employed. (Koc, 2025)

Research on Estonian language technologies demonstrates that pretrained language models, when combined with curated datasets, can significantly improve text simplification and accessibility for speakers of smaller languages (Barbu, Muru and Malva, 2025). However, such efforts remain resource-intensive and legally constrained, as the availability of suitable training data is often limited by copyright restrictions and licensing barriers.

Without structural intervention, digital language technologies are likely to continue favouring high-resource languages, even within multilingual political entities such as the European Union. Smaller EU languages face systemic disadvantages in digital infrastructure development, despite formal commitments to linguistic diversity (Muischnek, 2023). From this

perspective, the governance of text and data mining becomes indirectly linked to cultural policy objectives, as restrictive approaches to data use may inadvertently undermine efforts to sustain linguistic diversity in digital spaces.

Developing technologies for low-resource languages requires not only access to data but also careful consideration of cross-linguistic transfer effects. Studies on speech and text models for Finno-Ugric and other low-resource languages show that pretraining on high-resource languages can yield mixed outcomes and sometimes improve baseline performance while simultaneously introducing structural biases (Purason, Kuulmets and Fishel, 2025). These findings highlight the interaction between technical design choices and cultural factors in shaping how languages are represented and used within AI-mediated communication systems.

Ethical considerations and morality are equally relevant in this debate (Dutt, Chochia and Hoffmann, 2025). Ethical scholarship on language loss in the digital age cautions against treating AI as a neutral or automatic solution to linguistic decline. Shompa (2025) argues that while AI can support revitalisation efforts, it cannot substitute for social, educational, and institutional commitments to language use. Consequently, the role of LLMs in smaller language contexts should be understood as complementary rather than determinative, with these technologies embedded in broader cultural ecosystems rather than positioned as standalone preservation tools.

The cultural impact of LLMs is shaped not only by technological capacity but also by legal decisions governing data access and training practices. As LLMs depend heavily on extensive linguistic corpora, restrictions on the availability of training data can significantly influence the degree to which particular languages are represented within AI systems (Nazi, Hossain and Mamun, 2025). These disparities are particularly significant for smaller regional languages, where limited data availability is often compounded by legal uncertainty regarding the use of copyrighted materials for AI training. As a result, regulatory frameworks governing text and data mining may indirectly shape linguistic representation in AI technologies by determining which languages can realistically be incorporated into large-scale training datasets. In the case of languages such as Estonian, ensuring that legal and technological infrastructures support adequate data availability becomes an important factor in maintaining their functional presence within digital communication environments.

4.1 LLMs' possible effect on the Estonian language

The availability of sufficient digital linguistic material is essential for ensuring the continued functional use of smaller regional languages such as Estonian within modern technological environments. As communication, education, and access to information increasingly occur through digital platforms, languages that are insufficiently represented risk becoming marginalised in these domains. Modern language use and acquisition are closely connected to technological systems, which now serve as important environments for linguistic interaction and development (Melkonyan and Matevosyan, 2020). Ensuring that AI systems are trained on adequate Estonian-language data could therefore enable Estonian speakers to access and use digital tools more effectively in their native language, supporting both linguistic competence and the continued relevance of the Estonian language in digital communication.

Additionally, technological integration plays an important role in reinforcing linguistic presence and supporting cultural continuity. When languages are represented in digital systems, they can remain viable for use in education, communication, and knowledge distribution (Melkonyan and Matevosyan, 2020). If Estonian is insufficiently represented in AI systems, speakers may increasingly rely on dominant languages, potentially narrowing the functional scope of Estonian over time. Adequate representation within AI technologies would support its continued use across digital contexts, contributing to linguistic sustainability and helping Estonian to remain an active and functional language within evolving technological ecosystems.

A recent experiment in Estonia illustrates this concern. A report by *ERR Novaator*, examining the performance of AI systems in Estonian cultural and linguistic contexts, found that current models often struggle with culturally specific expressions and contextual nuances (Peetsalu, 2026).

The effectiveness of AI-generated content varies significantly across languages, largely due to disparities in the availability of training data. A study comparing English, Arabic, and Chinese revealed that AI-generated responses were consistently rated as less complete, actionable, and creative in non-English languages, and that human participants using AI-generated assistance in those languages produced lower quality work than English-speaking participants. Notably, these disadvantages were especially pronounced in technically complex tasks, where the lack of sufficient high-quality training data in non-English languages limited the usefulness of

AI-generated outputs (Koo, 2025). This finding suggests that insufficient representation in AI training datasets may reduce not only the quality of AI-generated content but also the effectiveness of human work supported by AI systems. Ensuring that LLMs are trained on adequate Estonian-language material would therefore be essential for maintaining linguistic equality and enabling Estonian speakers to benefit more fully from AI technologies.

The performance and usefulness of AI systems depend directly on the quantity and quality of linguistic data available for a given language. LLMs can achieve higher accuracy, contextual understanding, and generation quality when trained on extensive and diverse linguistic corpora. This disparity creates structural inequalities in AI-supported communication, as speakers of underrepresented languages may receive less accurate, less coherent, or less contextually appropriate outputs (Nazi, Hossain and Mamun, 2025). In the case of Estonian, ensuring sufficient representation within AI training datasets could therefore improve the ability of language models to process and generate Estonian accurately, enabling more effective digital communication, information access, and technological participation. This relationship between linguistic representation and communicative effectiveness underscores the broader principle that language development and use are shaped by interactions within technological environments.

Speech and language models trained on limited data from low-resource languages exhibit significantly reduced performance compared to those trained on high-resource languages, particularly in tasks requiring accurate linguistic representation and contextual understanding. Substantially increasing the volume and diversity of language-specific training data can improve model accuracy and linguistic reliability, enabling more effective speech recognition and language generation (Fong, Matassoni and Brutti, 2025).

Additionally, *Language Act (Estonia) 2011* reflects the importance of linguistic accessibility in public communication. The Act mandates foreign-language audiovisual works presented publicly to be accompanied by a corresponding Estonian translation. In practice, the translation of large volumes of digital and audiovisual content may increasingly rely on automated translation technologies, allowing these requirements to be met efficiently. For AI systems to fulfil this function effectively they must possess sufficient linguistic competence in Estonian, which in turn depends on adequate training data and technological development in the language.

4.2 Strengthening Estonian language representation

The above analysis demonstrates that the representation of smaller languages in LLMs depends not only on technological capacity but also on the legal frameworks governing access to linguistic data. The *Copyright Act (Estonia) 1992* contains provisions that enable text and data mining under certain conditions. Section 17^1 of the Act defines TDM broadly as an automated analytical method aimed at analysing text and data in digital form to generate information such as patterns, trends, and correlations. Sections 19^1 and 19^2 of the *Copyright Act (Estonia) 1992* further establish exceptions allowing the reproduction of lawfully accessible works for the purposes of text and data mining. In addition, Section 20 permits cultural heritage institutions to reproduce works in their collections for preservation and other non-commercial purposes. Taken together, these provisions provide a legal foundation upon which a more ambitious national strategy for AI training in the Estonian language could be developed.

One possible reform would involve establishing a publicly supported national LLM training infrastructure dedicated specifically to the Estonian language (Martínez-Ramil *et al*., 2023). Under such a model, the government of Estonia could create a publicly funded or a public–private partnership tasked with developing LLMs trained on Estonian-language corpora. The institution responsible for coordinating such activities could be the National Library of Estonia or another cultural heritage institution operating under the Ministry of Culture. Such an institutional structure could facilitate large-scale TDM of Estonian-language materials within a more clearly defined and stable legal environment. By transferring LLM training under a cultural heritage institution, the activity could be framed as scientific and cultural research rather than purely commercial data extraction. If large-scale Estonian-language AI training were conducted primarily through research organisations or cultural heritage institutions, it could potentially fall within the scope of the scientific research exception established in Article 3 of the DSM Directive.

This approach could help reduce disparities between low-resource and high-resource usage in AI infrastructures by supporting the continued visibility and functional use of the Estonian language. Importantly, such a proposal does not require abandoning existing copyright protections. Rather, it would involve clarifying and expanding the role of cultural heritage institutions within Estonia's copyright framework to facilitate responsible TDM for language technology development.

Similar discussions have emerged in Nordic countries. In Finland, legal scholars and policymakers have increasingly examined how the data mining exception in the Finnish *Copyright Act* can support the development of national-language AI systems. These discussions emphasise the importance of enabling access to copyrighted works for text and data mining purposes while maintaining safeguards for rightsholders (Somppi *et al*., 2025).

Norway has also begun exploring the development of national LLMs aimed at strengthening the digital presence of Norwegian languages. Public initiatives seek to combine governmental funding, research institutions, and private technology actors to develop domestic AI systems trained on Norwegian-language materials (CryptoPolitan, 2024). These initiatives illustrate how smaller language communities are increasingly pursuing coordinated national strategies to ensure linguistic representation in AI systems.

5. Conclusion

The rapid development of generative AI and LLMs has exposed both the strengths and limitations of existing copyright frameworks governing text and data mining.

Regarding the first research question, the current European Union framework is primarily defined by Articles 3 and 4 on copyright in the Digital Single Market Directive ('Council directive (EU) 2019/790', 2019), which establish exceptions allowing the reproduction of lawfully accessible works for text and data mining. Although these provisions were designed to create legal certainty for automated data analysis, emerging national case law indicates that their application to generative AI training remains contested.

The second research question concerned how the European Union approach compares with other jurisdictions. The analysis demonstrates that the EU has adopted a structured legislative model centred on specific TDM exceptions. In the United States, reliance on the flexible fair use doctrine allows courts to evaluate AI training practices on a case-by-case basis, but this approach also creates considerable uncertainty (Chesterman, 2025).

In contrast, several Asian jurisdictions have developed alternative regulatory models. Japan has introduced a regime allowing broad information analysis, which has been interpreted to facilitate AI training using copyrighted

materials. Singapore has adopted a statutory exception for computational data analysis that applies to both commercial and non-commercial uses but remains limited by lawful access requirements and restrictions on circumventing technological protection measures (Terui and Chen, 2025). China, by comparison, continues to rely largely on general copyright and unfair competition rules (Lu and Yuzhi, 2024, pp. 4–5).

The third research question addressed how existing regulatory frameworks can accommodate the technological realities of AI training without undermining copyright protection or deepening cultural and linguistic inequalities. The analysis suggests that achieving such a balance requires recognising that AI training practices depend heavily on large-scale linguistic datasets. For languages such as Estonian, this challenge may require proactive legal and institutional strategies, including facilitating text and data mining through cultural heritage institutions or developing publicly supported language technology initiatives.

Beyond these legal considerations, the findings of this study highlight that the regulation of text and data mining has broader implications for linguistic development and cultural sustainability. As theoretical perspectives suggest, technological systems increasingly shape how languages function within modern communication ecosystems (Eslit, 2023). Consequently, the legal rules governing access to training data influence not only technological innovation but also the future digital presence of smaller languages. As litigation and regulatory debates continue to evolve, the relationship between copyright law, artificial intelligence, and linguistic diversity will remain a critical area of legal and cultural policy.

Johanna Priske holds a master's degree in law from Tallinn University of Technology. She graduated with a bachelor's degree from the same institution in 2024. Her bachelor's thesis, *Copyright infringement in video games in the EU and the US*, examined comparative approaches to copyright protection in digital gaming environments. Johanna Priske has completed an internship at the Estonian Ministry of Justice, now Ministry of Justice and Digital Affairs, in the intellectual property and competition law division. Her research interests focus on intellectual property law, particularly copyright in digital environments. She is currently working as a junior lawyer, practising primarily in the field of debt law.

Dr **Pawan Kumar Dutt** is an associate professor of intellectual property law at TalTech Law School of Tallinn University of Technology. His educational background includes degrees in chemistry, law, and a PhD in management sciences. He is an advocate and solicitor and a member of various law associations. Dutt has practiced law since 1998. His areas of research are intellectual property law, digital law, and competition law, and he has published widely in these areas. He has been involved in several international projects concerning the study of this field, funded by the European Union, USAID, and others, both within and outside Europe.

References

Alonso, E. and Lucchi, N. (2025) 'AI and copyright "hallucinations": does the text and data mining exception really support generative AI training?', *European Intellectual Property Review*, 47(9), pp. 515–526.

Barbu, E., Muru, M.L. and Malva, S.M. (2025) 'Improving Estonian text simplification through pretrained language models and custom datasets', in G. Angelova, M. Kunilovskaja, M. Escribe and R. Mitkov (eds) *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing: Natural Language Processing in the Generative AI Era, Varna, 8–10 September*, pp. 133–142. Available at: https://doi.org/10.26615/978-954-452-098-4-016

Bergin, R. (no date) *Media richness theory*. Center for Homeland Defense and Security, Department of National Security Affairs, Naval Postgraduate School.

Chesterman, S. (2025) 'Good models borrow, great models steal: intellectual property rights and generative AI', *Policy and Society*, 44(1), pp. 23–37. Available at: https://doi.org/10.1093/polsoc/puae006

Copyright Act (Estonia) 1992. Autoriõiguse seadus (AutÕS). RT I 1992, 49, 615, entry into force 12 December.

Copyright Act (Japan) 1970. Law No. 48, promulgated on May 6.

Copyright Act (Singapore) 2021. Entry into force 21 November.

Copyright Law (China) 2010. Order of the President of the People's Republic of China, No. 26, entry into force 1 April.

'Council directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC' (2019) *Official Journal* L 130, pp. 92–125. Available at: https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng (Accessed: 10 March 2026).

'Council regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU)

No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)' (2019) *Official Journal* L 2024/1689, 12 July.

CryptoPolitan (2024) 'Norway is on the way to advance its LLM game', *Cryptorank*, 14 April. Available at: https://cryptorank.io/news/feed/e07e8-norway-is-on-the-way-to-advance-its-llm-game (Accessed: 10 March 2026).

Crystal, D. (2006) *Language and the internet*. 2nd edn. Cambridge University Press.

Daft, R.L. and Lengel, R.H. (1986) 'Organizational information requirements, media richness and structural design', *Management Science*, 32(5), pp. 554–571. Available at: https://doi.org/10.1287/mnsc.32.5.554

District Court for the Northern District of California (2025a) 'Bartz *et al.* v. Anthropic PBC', Case No. C 24-05417 WHA. Decision, 23 June.

District Court for the Northern District of California (2025b) 'Kadrey *et al.* v. Meta Platforms, Inc.', Case No. 23-cv-03417-VC. Decision, 25 June.

District Court of Amsterdam (2024) Case No. C/13/737170 / HA ZA 23-690; ECLI:NL:RBAMS:2024:6563. Decision of 30 October.

Dornis, T.W. (2025) 'The training of generative AI is not text and data mining', *European Intellectual Property Review*, 47(2), pp. 65–78.

Dusollier, S. (2020) 'The 2019 Directive on copyright in the Digital Single Market: some progress, a few bad choices, and an overall failed ambition', *Common Market Law Review*, 57(4), pp. 979–1030. Available at: https://doi.org/10.54648/COLA2020714

Dutt, P.K., Chochia, A. and Hoffmann, T. (2025) 'Understanding the manifold approaches that shape research, development and innovation policies', *TalTech Journal of European Studies*, 15(3), pp. 3–23. Available at: https://doi.org/10.2478/bjes-2025-0029

Eslit, E.R. (2023) 'Neo-language acquisition and learning theory: a new perspective on language teaching', *Preprints*, pp. 1–10. Available at: https://doi.org/10.20944/preprints202309.1905.v1

Fong, S., Matassoni, M. and Brutti, A. (2025) 'Speech LLMs in low-resource scenarios: data volume requirements and the impact of pretraining on high-resource languages', *Proc. Interspeech 2025*. Available at: https://doi.org/10.21437/Interspeech.2025-764

Ge, C. (2025) 'Beyond fair use and opt-out: forging a hybrid copyright path for generative AI in China', *Innovative Applications of AI*, 2(3), pp. 141–146. Available at: https://doi.org/10.70695/10.70695/IAAI202503A9

Guadamuz, A. (2025) 'The EU's Artificial Intelligence Act and copyright', *Journal of World Intellectual Property*, 28(1), pp. 213–219. Available at: https://doi.org/10.1111/jwip.12330

Haokip, T. (2022) 'Artificial intelligence and endangered languages', *Journal of North East India Studies*, 12(1), pp. 1–7. Available at: https://doi.org/10.2139/ssrn.4212504

Hugenholtz, P.B. and Senftleben, M. (2011) 'Fair use in Europe: in search of flexibilities', *SSRN*, pp. 1–30. Available at: https://doi.org/10.2139/ssrn.1959554

Koc, V. (2025) Generative AI and large language models in language preservation: opportunities and challenges. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2501.11496

Koo, W.W. (2025) 'Cross-lingual effects of AI-generated content on human work', *Scientific Reports*, 15, article 30949. Available at: https://doi.org/10.1038/s41598-025-16650-w

Language Act (Estonia) 2011. Keeleseadus (KeeleS). RT I, 18 March, entry into force 1 July, 1.

Leistner, M. and Antoine, L. (2025) 'TDM and AI training in the European Union – from '*LAION*' to possible ways ahead?', *GRUR International*, 74(11), pp. 1027–1044. Available at: https://doi.org/10.1093/grurint/ikaf114

Lu, W. and Yuzhi, W. (2024) 'Comparative study of TDM technology systems from an international perspective', in *9th International Conference on Computer and Communication Systems (ICCCS), Xi'an, China*, pp. 1172–1177. Available at: https://doi.org/10.1109/ICCCS61882.2024.10603045

Manteghi, M. (2025) 'Kneschke vs. LAION: opening fresh perspectives on AI training and TDM exceptions', *European Law Blog*, pp. 1–5. Available at: https://doi.org/10.21428/9885764c.5f1ce2dd

Margoni, T. and Kretschmer, M. (2022) 'A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology', *GRUR International*, 71(8), pp. 685–701. Available at: https://doi.org/10.2139/ssrn.3886695

Martínez-Ramil, P., Bolaños-Frasquet, H., Hamuľák, O. and Kerikmäe, T. (2023) 'A multidimensional understanding of EU's digital sovereignty', in D. Ramiro Troitiño, T. Kerikmäe and O. Hamuľák (eds) *Digital development of the European Union: an interdisciplinary perspective*. Springer International Publishing, pp. 235–247. Available at: https://doi.org/10.1007/978-3-031-27312-4_15

Melkonyan, A. and Matevosyan, A. (2020) 'Technology-assisted foreign language learning (TALL) in the digital age', in *SHS Web of Conferences*, 88, pp. 1–7. Available at: https://doi.org/10.1051/shsconf/20208802005

Muischnek, K. (2023) 'Language report Estonian', in G. Rehm and A. Way (eds) *European language equality*. *Cognitive technologies*. Springer, pp. 131–134. Available at: https://doi.org/10.1007/978-3-031-28819-7_14

Municipal Court of Appeals of Budapest (2024) Case No. 9.Pf.20.353/2024/6-II. Decision of 3 December.

Nazi, Z.A., Hossain, M.R. and Mamun, F.A. (2025) 'Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting', *Natural Language Processing Journal*, 10, pp. 1–12. Available at: https://doi.org/10.1016/j.nlp.2024.100124

Ong, S. (2025) 'Beyond the exception: licensing, access, and the realities of text and data mining in the US, UK, and Singapore', *SSRN*, pp. 1–22. Available at: https://doi.org/10.2139/ssrn.5439658

Peetsalu, J. (2026) '"AK Nädal" uuris, mida mõistab tehisaru eesti keelest ja kultuurist' [*The weekly news investigated what AI understands about Estonian language and culture*], *ERR Novaator*. Available at: https://novaator.err.ee/1609961504/ak-nadal-uuris-mida-moistab-tehisaru-eesti-keelest-ja-kultuurist (Accessed: 10 March 2026).

Purason, T., Kuulmets, H.A. and Fishel, M. (2025) 'LLMs for extremely low-resource Finno-Ugric languages', in L. Chiruzzo, A. Ritter and L. Wang (eds) *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6692–6712. Available at: https://doi.org/10.18653/v1/2025.findings-naacl.373

Quintais, J.P. (2025) 'Generative AI, copyright and the AI Act', *Computer Law & Security Review*, 56, article 106107, pp. 1–17. Available at: https://doi.org/10.1016/j.clsr.2025.106107

Radeisen, A. (2026) 'Open foundation models and TDM exceptions to copyright—building blocks for an AI ecosystem', *GRUR International*, 75(3), pp. 224–234. Available at: https://doi.org/10.1093/grurint/ikag002

Regional Court of Hamburg (2024) Case No. 310 O 227/23. Decision, 27 September.

Shamov, O. (2025) 'AI and copyright: from a doctrinal crisis to a hybrid model of collective licensing', *Legal Horizons*, 26(3), pp. 68–76. Available at: https://doi.org/10.54477/LH.25192353.2025.3.pp.68-76

Shompa, S.N. (2025) 'Language death in the digital age: can AI save endangered tongues?' *International Journal of Ethical AI Application*, 1(1), pp. 33–42. Available at: https://doi.org/10.64229/ymjyqc91

Short, J., Williams, E. and Christie, B. (1976) *The social psychology of telecommunications*. Wiley.

Shortis, T. (2007) 'Gr8 txtpectations: the creativity of text spelling', *English Drama Media*, (8), pp. 21–26. Available at: https://nl.ijs.si/janes/wp-content/uploads/2014/09/shortis07.pdf (Accessed: 10 March 2026).

Söderlund, K. and Larsson, S. (2024) 'Enforcement design patterns in EU law: an analysis of the AI Act', *Digital Society*, 3, article 41, pp. 1–21. Available at: https://doi.org/10.1007/s44206-024-00129-8

Somppi, A., Mäkitalo-Saarinen, H., Siitonen, P. and Schröder, V. (2025) 'Using copyright-protected works to train AI from the perspective of the data mining exception of the Finnish Copyright Act', Hannes Snellman. Available at: https://

www.hannessnellman.com/news-and-views/blog/using-copyright-protected-works-to-train-ai-from-the-perspective-of-the-data-mining-exception-of-the-finnish-copyright-act/ (Accessed: 10 March 2026).

Terui, T. and Chen, J. (2025) 'Generative AI, copyright, and cultural policy: understanding the mechanisms behind Japan and Singapore's AI-friendly approaches', *International Journal of Cultural Policy*, pp. 1–19. Available at: https://doi.org/10.1080/10286632.2025.2588195

Thongmeensuk, S. (2024) 'Rethinking copyright exceptions in the era of generative AI: balancing innovation and intellectual property protection', *Journal of World Intellectual Property*, 27(2), pp. 278–295. Available at: https://doi.org/10.1111/jwip.12301

United States Copyright Act 1976, 17 U.S.C. § 107.

Van Lier, L. (2004) *The ecology and semiotics of language learning: a sociocultural perspective*. Kluwer Academic Publishers.

Vardanyan, L., Kocharyan, H., Hamuľák, O., Gábriš, T. and Martínez-Ramil, P. (2026) 'Balancing monopolisation and innovation in the esports industry: searching for *aurea mediocritas* between public and private interests in intellectual property rights', *Law, Innovation and Technology*, 18(1), pp. 188–232. Available at: https://doi.org/10.1080/17579961.2026.2633686

Walther, J.B. (2011) 'Theories of computer-mediated communication and interpersonal relations', *ResearchGate*, pp. 443–479.

Wang, Z., Chu, Z., Doan, T.V., Ni, S., Yang, M. and Zhang, W. (2025) 'History, development, and principles of large language models: an introductory survey', *AI and Ethics*, 5(3), pp. 1955–1971. Available at: https://doi.org/10.1007/s43681-024-00583-7

Wrench, J.S., Punyanunt-Carter, N.M. and Thweatt, K.S. (2026) *Interpersonal communication: a mindful approach to relationships*. LibreTexts.