10

Large Language Models

Michal Ovádek

10.1 INTRODUCTION

The empirical study of European Union (EU) law has evolved significantly in the last few decades, driven by both advances in theoretical understanding and methodological innovations.[1] Traditionally, this field has relied on doctrinal analyses of case law, legislative texts, and policy documents to draw insights into the functioning and development of EU law. However, the emergence of large language models (LLMs) presents a novel and potentially transformative tool for scholars in this domain. This chapter examines the integration of LLMs into empirical legal studies of EU law, focusing on some of their capabilities, associated challenges, and future prospects.

Large language models are a type of machine learning algorithm (or artificial intelligence) that processes and most commonly generates human-like text based on patterns learned from vast amounts of data. These models, most famously exemplified by technologies such as OpenAI's Generative Pre-trained Transformer (GPT) series, have demonstrated remarkable capabilities in generating coherent and contextually appropriate text outputs. For legal scholars, the allure of LLMs lies in their ability to digest and analyse large corpora of legal texts much more swiftly and comprehensively than human researchers could. This capability not only enhances the effesseliciency of research but also opens up new avenues for hypothesis testing and data-driven insights.

Employing LLMs in the study of EU law can help uncover latent patterns and trends across numerous legal documents, including judgments of the Court of

[1] Arthur Dyevre, 'EU Judicial Behaviour Research: A Look Back and a Look Ahead' (2024) 25(3) *European Politics and Society* 468; Gareth Davies, 'The Relationship between Empirical Legal Studies and Doctrinal Legal Research' (2020) 13(2) *Erasmus Law Review* 3–12; Arthur Dyevre, Wessel Wijtvliet, and Nicolas Lampach, 'The Future of European Legal Scholarship: Empirical Jurisprudence' (2019) 26(3) *Maastricht Journal of European and Comparative Law* 348.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

Justice of the European Union (CJEU) and EU legislation. By analysing these texts, LLMs can assist scholars in identifying shifts in legal reasoning, changes in policy focus, and the evolution of legal norms within the EU. Moreover, these models can facilitate comparative analysis, helping legal researchers to draw parallels or highlight differences between various jurisdictions within the EU, a task that is particularly challenging given the multilingual nature of law in the EU. While multilingualism is not a unique property of EU law, it sets it apart from most domestic jurisdictions and as such poses an additional set of challenges for the study of EU law.[2]

However, the application of LLMs in legal studies is not without challenges. One significant concern is the accuracy and reliability of the outputs generated by these models. While LLMs are proficient at mimicking the structure and style of legal texts, their understanding of complex legal concepts and the nuances of legal reasoning can be superficial. This limitation raises questions about the validity of conclusions drawn solely from LLM-generated analyses. Therefore, it is imperative for legal scholars to critically assess the outputs of LLMs and, where necessary, complement them with different methods. Furthermore, ethical considerations must be addressed when integrating LLMs into legal research. Issues such as data privacy, consent, and the potential for algorithmic bias are particularly pertinent in the legal domain. These models are trained on existing legal documents, and thus, they may perpetuate existing biases found in the source material. Legal scholars must be vigilant about these risks and work towards developing methodologies that mitigate bias and uphold the ethical standards of legal research.

This chapter provides an introduction to LLMs and highlights the key challenges of using them in empirical legal research. The next section explains in non-technical terms what LLMs are and how they differ from smaller language models. Subsequently, I look at perhaps the most common use-case of LLMs in legal research, namely information retrieval. I then discuss the practical costs and benefits of using LLMs relative to smaller models before turning to what I see as the most important aspect of deploying LLMs in research – validating their outputs and how to use them in downstream empirical analysis.

10.2 WHAT ARE LLMS?

LLMs represent a critical development of natural language processing technologies, a key subset of machine learning and artificial intelligence aimed at enabling

[2] Elina Paunio, *Legal Certainty in Multilingual EU Law: Language, Discourse and Reasoning at the European Court of Justice* (Routledge 2016); Aleksandra Čavoški, 'Interaction of Law and Language in the EU: Challenges of Translating in Multilingual Environment' (2017) 27 *Journal of Specialised Translation* 58.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

machines to understand and generate human language.[3] These models differ from earlier language models not only in scale but also in their capacity to handle a broader array of linguistic tasks due to their architecture and the sheer volume of data they are trained on.[4]

At their core LLMs utilise the transformer architecture, which is pivotal for handling sequential data – in this case words – more effectively than previous models that relied on recurrent neural networks (RNNs) or convolutional neural networks (CNNs). The transformer uses an attention mechanism that allows the model to weigh the importance of different words in a sentence or passage dynamically, regardless of their position. This mechanism enables the model to generate more contextually appropriate outputs, maintaining coherence over longer stretches of text.[5]

The training process for LLMs involves extensive pre-training on diverse datasets comprising a wide range of internet-sourced texts, including books, articles, and websites.[6] This phase is generally unsupervised, with the model learning to predict the next word in a sequence without direct task-specific guidance. Such extensive pre-training helps the model develop a generalised understanding of language, grammar, and context. Fine-tuning usually follows, where the LLM is trained on more specific data, enhancing its capabilities for particular tasks through supervised learning, where it adjusts based on feedback tied to specific objectives.

What sets LLMs apart from smaller language models that also use the transformer architecture – such as the Bidirectional Encoder Representations from Transformers (BERT) class of models – is primarily their scale. LLMs such as GPT-4 are characterised by a much larger number of parameters, often in the hundreds of billions. This scale allows for a more nuanced understanding of language and enables the models to perform a wider variety of tasks with higher accuracy and less input from users once training is complete.

Moreover, the volume and diversity of the training data for LLMs exceed what smaller transformer-based models typically process.[7] This extensive training allows LLMs to generalise across more contexts, making them versatile tools for applications that require understanding and generating human-like text across diverse

[3] Idan A. Blank, 'What Are Large Language Models Supposed to Model?' (2023) 27(11) *Trends in Cognitive Sciences* 987.

[4] Jingfeng Yang, Hongye Jin, and Ruixiang Tang et al., 'Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond' (2024) 18(6) *ACM Transactions on Knowledge Discovery from Data* 1.

[5] Ashish Vaswani, Noam Shazeer, and Niki Parmar et al., 'Attention Is All You Need' (2017) Advances in Neural Information Processing Systems.

[6] Tom B. Brown, Benjamin Mann, and Nick Ryder et al., 'Language Models Are Few-Shot Learners' (2020) arXiv preprint arXiv:2005.14165.

[7] The open-access or proprietary nature of the training data is a contentious issue. Even many nominally open-source models such as Llama do not release their training data. See Hugo Touvron, Thibaut Lavril, and Gautier Izacard et al., 'Llama: Open and Efficient Foundation Language Models' (2023) arXiv preprint arXiv:2302.13971.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

domains and styles. This capability is critical in areas like legal analysis, where the ability to parse through and synthesise vast amounts of complex textual information is necessary. All of this comes at a much higher computational – and by extension economic and environmental – cost.

In addition, the LLM paradigm has come with natural language control and zero- to few-shot learning. When fine-tuning BERT models, users typically use structured datasets with many examples of what the model is meant to learn, for example a label that should be associated with certain patterns in text. When fine-tuning or tasking LLMs, users can instead rely on natural language prompts which the model is capable of interpreting as instructions. And unlike BERT-sized models, LLMs tend to perform well on many tasks with only the content of the natural language prompt without needing to first learn from a large, task-specific dataset (although mileage will vary depending on the quality of the model and difficulty of the task).

As we shall see, however, the creation of LLMs has not fully crowded out smaller – especially BERT-like – models. On the contrary, the artificial intelligence research community has been looking for ways to leverage the power of LLMs in conjunction with smaller models as a means of reducing computational cost, increasing speed, and in some cases even improving overall performance.

10.3 INFORMATION RETRIEVAL

Perhaps the most obvious task in which LLMs can assist EU lawyers is legal information retrieval (LIR).[8] LIR encompasses the processes and methodologies employed to locate and extract relevant legal information, such as statutes, regulations, and court decisions, from vast and potentially complex legal databases. It is a critical aspect of legal research and practice, enabling legal professionals to access precedents, interpret laws, and formulate legal arguments with precision and efficiency. The importance of LIR lies in its ability to facilitate informed decision-making and ensure that legal practitioners can swiftly and accurately navigate the law. However, achieving effective LIR has historically posed significant challenges due to the intricate and often arcane nature of legal language, the vast volume of legal documents, and the necessity for contextual understanding. Traditional search techniques frequently fall short in addressing these complexities, leading to incomplete or imprecise results. Even state-of-the-art pre-LLM methods have produced relatively paltry retrieval accuracy.[9]

[8] Ahmed Izzidien, Holli Sargeant, and Felix Steffek, 'LLM vs. Lawyers: Identifying a Subset of Summary Judgments in a Large UK Case Law Dataset' (2024) arXiv preprint arXiv:2403.04791.

[9] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka, 'Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023' in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (2023) 472; Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu, 'THUIR@COLIEE 2023: Incorporating Structural Knowledge

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

More specifically, a core challenge in legal research is the extraction of legally relevant information – such as the identity of the applicant and legal claims being made – from court decisions. Traditionally, this is a laborious manual process whereby a legal expert or assistant reads through the decisions and notes down key information. Scaling it up to thousands of decisions is prohibitively expensive, especially for academic researchers. As shown in a recent study on extracting legal information from UK employment tribunal decisions, LLMs might be capable of performing this task very well.[10] Compared with human annotations of the same decisions, the authors adjudged the LLM to have extracted the correct information between 90 and 100 per cent of the time, depending on what specifically was being extracted.

I showcase the method on an example from EU law. I only very slightly simplify the prompt developed by de Faria et al. for this purpose and apply it to the well-known *Portuguese Judges* judgment concerning judicial independence, handed down by the CJEU in 2018.[11] I prompt GPT-4o in the following way:

> You are a legal assistant. Your task is to read through the court decisions which I will send you, and extract the following information for each input: 1. facts of the case of the specific court decision; 2. claims made in the specific court decision and considered in the specific court decision. Do not include any claim which has already been decided in any previous decision; 3. any references to legal statutes, acts, regulations, provisions and rules, including the specific number(s), section(s) and article(s) of each of them, and including procedural rules; 4. references to precedents and other court decisions; 5. general case outcome; 6. general case outcome summarised using one of the following four labels – 'claimant wins', 'claimant loses', 'claimant partly wins' and 'other'. Note that the label 'other' is to be reserved for situations in which the result cannot be determined or where the outcome cannot be described in terms of winning or losing.
>
> <text of the ruling>

We receive the following response from the LLM:

> 1. Facts of the Case:
>
> The Associação Sindical dos Juízes Portugueses (ASJP), representing members of the Tribunal de Contas (Court of Auditors, Portugal), challenged temporary salary reductions imposed on the court's members. These reductions were part of the Portuguese State's budgetary policy aimed at reducing the excessive budget deficit, based on Law No 75/2014. The ASJP argued that these

into Pre-Trained Language Models for Legal Case Retrieval' (2023) arXiv preprint arXiv:2305.06812.

[10] Joana Ribeiro de Faria, Huiyuan Xie, and Felix Steffek, 'Automatic Information Extraction From Employment Tribunal Judgements Using Large Language Models' (2024) arXiv preprint arXiv:2403.12936.

[11] Case C-64/16 *Associação Sindical dos Juízes Portugueses* ECLI:EU:C:2018:117.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

reductions violated the principle of judicial independence enshrined in the Portuguese Constitution and EU law.

2. Claims:
   The primary claim considered in this decision was whether the principle of judicial independence under the second subparagraph of Article 19(1) TEU and Article 47 of the Charter of Fundamental Rights of the European Union precludes the implementation of salary reduction measures on members of the judiciary, such as those imposed by Law No 75/2014.
3. References to Legal Statutes, Acts, Regulations, Provisions, and Rules:
   Article 19(1) TEU: Pertaining to judicial independence and effective judicial protection.
   Article 47 of the Charter: Right to an effective remedy and a fair trial.
   Portuguese Law No 75/2014: Mechanisms for the temporary reduction of remuneration in the public sector.
   Portuguese Law No 159-A/2015: Gradual elimination of remuneration reduction measures.
   Lei n.º 98/97: Organization and procedure of the Tribunal de Contas.
4. References to Precedents and Other Court Decisions:
   Judgment of 4 May 2016, Philip Morris Brands and Others (C-547/14): Relevance of the reasons for interpreting EU law.
   Judgment of 21 December 2016, Associazione Italia Nostra Onlus (C-444/15): Grounds for refusal to rule on a question.
   Judgment of 3 October 2013, Inuit Tapiriit Kanatami and Others v Parliament and Council (C-583/11 P): Effective judicial protection and rule of law.
   Judgment of 28 April 2015, T & L Sugars and Sidul Açúcares v Commission (C-456/13 P): Judicial review and compliance with EU law.
   Judgment of 16 February 2017, Margarit Panicello (C-503/15): Criteria for determining a 'court or tribunal' under EU law.
   Judgment of 26 May 2016, Județul Neamț and Județul Bacău (C-260/14 and C-261/14): Application or interpretation of EU law by national courts.
   Judgment of 19 September 2006, Wilson (C-506/04): Protection of judicial independence.
5. General Case Outcome:
   The Court of Justice of the European Union (CJEU) ruled that the principle of judicial independence under EU law does not preclude general salary reduction measures, such as those implemented by Law No 75/2014, from being applied to the members of the judiciary, provided that the measures are general in nature, not specifically aimed at the judiciary, and are justified by overriding reasons of public interest.
6. General Case Outcome Label:
   Claimant loses

The extracted information is largely correct. In simplified terms, ASJP did indeed lose, as the CJEU legitimised the Portuguese government's austerity measures.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

The main issue for the model is identifying all court citations (although it also missed Article 2 TEU). It missed references to six out of thirteen rulings cited by the CJEU, probably also because it did not recognise Opinions (1/09 and 2/13) as decisions. This indicates worse performance than that reported by de Faria et al., but it should be noted that I made few adaptations to their prompt which was designed to extract information from a rather different legal context. It is likely that EU law-specific prompt engineering would yield better results, as would additional fine-tuning on EU law texts.[12]

At the same time, we do not know whether the ASJP decision was included in GPT-4o's training data, but as CJEU rulings are publicly available, this is likely. As such, the LLM would be likely to perform worse on texts which it had not seen prior. The extent to which this constitutes a problem for the researcher-user depends on their specific task and data, but it is certainly something one should be cognisant of in the process of working with LLMs.

Furthermore, it is worth noting the breadth of information the LLM is capable of extracting from the text using only a simple prompt, which speaks to the model's versatility. There is in fact no other technology that would be able to reliably retrieve certain kinds of information from a large unstructured text input, such as what claims ASJP made in the legal proceedings. However, as the omissions in the citations indicate, LLMs can struggle with highly factual tasks that require a lot of precision. As a result, computer scientists developed a workflow called retrieval-augmented generation (RAG) which combines LLMs' generative capabilities with access to factual information.[13]

How can RAGs help with our citation extraction problem? We can create a database of CJEU court citations[14] – for example using a smaller but specialised named entity recognition (NER) model – and connect the LLM to it. The level of sophistication involved in linking the user's natural language prompts, the LLMs generative process, and access to the citation database will vary depending on how the data is structured. The simplest version of RAGs uses (cosine) similarity between the prompt and the documents stored in the database to retrieve the most relevant

[12] Michal Ovádek, Joshua Fjelstul, Daniel Naurin, and Johan Lindholm, 2023. 'The IUROPA Text Corpus' in Johan Lindholm, Daniel Naurin, and Urška Šadl et al., The Court of Justice of the European Union (CJEU) Database, IUROPA, https://iuropa.pol.gu.se/.

[13] Patrick Lewis, Ethan Perez, and Aleksandra Piktus et al., 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks' (2020) 33 *Advances in Neural Information Processing Systems* 9459.

[14] One could reuse the metadata from Eur-Lex for this purpose, but they are not complete and contain a considerable number of mistakes. See Michal Ovádek, 'A Note of Caution on CJEU Databases' (2024) 3(2) *European Law Open* 353–59; Joshua Fjelstul, Daniel Naurin, Johan Lindholm, and Michal Ovádek, 'Theorizing and Measuring the Precedential Value of Case Law: Lessons from the European Union' (2024) Working Paper.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

entries to 'augment' the LLM's generative response. A more performant but also technically involved solution would entail translating the user's prompt into a structured-query language prompt which identifies exactly the necessary information in the database.[15] The individual-database RAG idea can be generalised into a broader notion of agentic workflows whereby the process of generating an answer to a prompt involves decomposing the problem into smaller tasks which are subsequently tackled with greater accuracy by specialised machine learning models (known as agents).[16] From the user's perspective, a great deal of the attraction of this system is that it can still be controlled using natural language instructions, showing how LLMs also contribute to bridging the skill gap between researchers with (most quantitative social scientists) and without (most legal scholars) programming skills.

10.4 SCALING UP, SCALING DOWN

At the time of writing, it is generally true that LLMs are slower and much more expensive to train and run than smaller models. Training GPT-4, the state-of-the-art model as of 2024, cost upwards of 100 million USD.[17] Its faster iteration, GPT-4o is priced by OpenAI as of May 2024 at 5 USD per 1 million input tokens and 15 USD for every 1 million output tokens. For illustration, the full prompt from our example above had 6,405 tokens (cost: 0.032 USD), while GPT's response was 672 tokens long (cost: 0.01 USD). The ASJP ruling is on the shorter end of the spectrum, however, so let's assume an average CJEU decision to cost at least 5 cents USD to process in the way shown above. Given that there are approximately 48,000 decisions (in French) across the General Court and the Court of Justice, scaling up the processing would cost at least ~ 2,500 USD. In other words, an LLM can under some conditions parse the entire CJEU corpus for the equivalent of a month's salary of a research assistant.

For larger or more complex tasks, the costs can become too high – and environmental costs are even higher. However, and in any case, we might be interested in models that produce simpler outputs faster, and LLMs can still be helpful. For example, let's say we are interested in finding out whether the CJEU agrees or disagrees (or neither) with the Commission in each paragraph. Instead of feeding an

[15] Jinyang Li, Binyuan Hui, and Ge Qu et al., 'Can LLM Already Serve as a Database Interface? A Big Bench for Large-Scale Database Grounded Text-to-SQLs' (2024) 36 *Advances in Neural Information Processing Systems* 42330.

[16] See, e.g., Bing Wang, Changyu Ren, and Jian Yang et al., 'MAC-SQL: Multi-Agent Collaboration for Text-to-SQL' (2023) arXiv preprint arXiv:2312.11242.

[17] Will Knight, 'OpenAI's CEO Says the Age of Giant AI Models Is Already Over' (*Wired*, 17 April 2023), https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

LLM millions of paragraphs, we can ask it to answer the question for a sample of a few thousand paragraphs and subsequently use this 'synthetic' dataset to train a BERT model.[18] This is the equivalent of asking a research assistant – or several, we can ask an LLM for multiple evaluations – to label the data to subsequently use for model training, but significantly more cheaply. The performance of the smaller model can be as high as that of the LLM, although results will vary depending on the complexity of the specific task. The fine-tuned BERT model will operate faster and at a fraction of the cost (both environmental and financial) than the LLM.

'Scaling down' or distilling the LLM knowledge into smaller models is going to work best for relatively straightforward tasks, such as text classification (labelling), which leverage more the LLM's superior understanding of text patterns than its generative capabilities. Similarly, a core limitation of this approach is the size of the context – how much text the LLM is able to 'read' when answering a query – necessary to make good inferences about the text. In our proposed task, the information the model needs is normally present within a single paragraph (the unit of analysis). We are relying on the assumption that most of the time when the CJEU discusses the arguments of the parties or interveners, such as the Commission, the identity of the actor and the Court's own position are present within the same paragraph. However, in more complex tasks, a paragraph or another relatively short unit of text might not contain all the necessary information, such as when the Court refers to an argument or precedent discussed in a previous paragraph. In such cases, the larger context size which LLMs work with (and smaller models do not) is likely to be critical to performance.

10.5 VALIDATION, VALIDATION, VALIDATION

I have so far at most alluded to a core challenge of doing research with LLMs. How do we know whether the LLM output is useful? Tackling this question appropriately is perhaps the most important role of applied researchers. The consensus in computational social science for the time being considers expert, human evaluation as the gold standard by which to judge any machine-generated output, including that of LLMs. Despite the fact that defining the gold standard is frequently a fraught affair – as demonstrated by plenty of research on coder reliability[19] – machine outputs ultimately serve the ends of human-led research and not vice versa (at least for

[18] https://huggingface.co/blog/synthetic-data-save-costs. Naoki Egami, Musashi Jacobs-Harukawa, Brandon M. Stewart, and Hanying Wei, 'Using Large Language Model Annotations for Valid Downstream Statistical Inference in Social Science: Design-Based Semi-Supervised Learning' (2023) arXiv preprint arXiv:2306.04746. call the labels produced by an LLM 'surrogate'.

[19] Slava Mikhaylov, Michael Laver, and Kenneth R. Benoit, 'Coder Reliability and Misclassification in the Human Coding of Party Manifestos' (2012) 20(1) *Political Analysis* 78; K. Krippendorf, *Content Analysis: An Introduction to Its Methodology* 4th edn. (Sage 2018); Kyle L. Marquardt and Daniel Pemstein, 'IRT Models for Expert-Coded Panel Data' (2018) 26(4) *Political Analysis* 431. See also Kim and Schroeder, Chapter 9 in this volume.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

now). In practice, this means that LLM output requires human – frequently expert – validation.

The main difference between validating LLM output compared to previous generations of text models is that our baseline expectations about the LLM's performance are considerably higher and generally more transferable. Several studies have shown that LLMs are capable of outperforming human annotators on certain labelling and sentiment detection tasks.[20] This makes sense – machines have the advantage of not suffering from inattention and fatigue and can therefore produce more consistent output than humans. Nonetheless, and somewhat paradoxically, the only way we can evaluate the performance of the models is by agreeing on a gold standard of what constitutes the ground truth. It requires deciding whether, for example, in the following paragraph the CJEU states that a rule of EU law has direct effect:

> Whilst it is true, as observed by the Italian and Danish Governments, that a directive cannot of itself impose obligations on an individual and cannot therefore be relied on as such against an individual (see Case C-91/92 Faccini Dori [1994] ECR I-3325, paragraph 20), that case-law does not apply where non-compliance with Article 8 or Article 9 of Directive 83/189, which constitutes a substantial procedural defect, renders a technical regulation adopted in breach of either of those articles inapplicable.

Human coders in the IUROPA project believed the answer is 'yes'.[21] GPT-4o instructed with the following prompt believes the answer is 'no':

> You are an expert in European Union law. Your sole task is to classify paragraphs of CJEU judgments. Specifically, you must decide whether the Court of Justice states in the paragraph that a rule of EU law has direct effect. Before making the choice [yes/no], read the paragraph carefully and reflect on the reasons that lead you to either conclusion. Answer with your reasoning and the choice made.

Where might the difference between human and GPT judgment arise from? The most likely culprit for the model's divergent interpretation is likely to be found in the overall tendencies in its training data. The model is presented with a difficult and rather marginal example of direct effect (incidental direct effect). There will be very few positive examples of this type of direct effect discussed in the training data. The

[20] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli, 'ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks' (2023) 120 *Proceedings of the National Academy of Sciences* e2305016120; Caleb Ziems, Willaim Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang, 'Can Large Language Models Transform Computational Social Science?' (2024) 50(1) *Computational Linguistics* 237.

[21] Olof Larsson, Johan Lindholm, and Daniel Naurin et al., 2022. 'The IUROPA CJEU Database: Issues and Positions' in Johan Lindholm, Daniel Naurin, and Urška Šadl et al., *The IUROPA Court of Justice of the European Union (CJEU) Database*, IUROPA, https://www.iuropa.pol.gu.se/.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

LLM therefore likely veers off towards its more general knowledge of the doctrine, which does not deal with or attribute direct effect to incidental situations.

There are no guarantees that the LLM will perform well on every task and there is no way of knowing this a priori. The machine will struggle, similarly to humans, on relatively more difficult tasks. Human disagreement is in principle an indicator for how well we should expect an LLM to perform a task and answer legal questions. Nonetheless, the manner and extent to which difficulty translates from humans to LLMs is going to be context-dependent. The existence of LLMs will invite constant experimentation to identify (difficult) tasks that the machine will do well (overperform or at least not significantly underperform humans). There is no theoretical upper bound on task difficulty that could not be tackled by the most powerful LLMs. In due time and with enough resources – including very large context windows enabling the feeding of thousands of pages of text – the best models should be able to, for example, analyse judicial doctrine as well as humans in a fraction of the time. In general, reasoning models are likely to perform better at complex classification tasks than non-reasoning models.

Establishing a gold-standard benchmark for the evaluation of LLMs requires following the same procedures as creating any human-coded dataset, that is, defining and delineating concepts and preparing a codebook explaining how labels relate to observations (input text) (see Chapter 9, Kim and Schroeder in this volume).[22] In some cases, the creation of the gold-standard dataset is uncontroversial, as with validating which precedents were cited by the CJEU in a decision. In others, coding agreement can be more difficult to find. Depending on the distribution of the underlying concepts, different sampling strategies might be appropriate. Frequently, we are working with imbalanced data in EU law. For example, there are many more cases without a direct effect angle than those with, and it might be more important to us theoretically to evaluate the positive class. Sampling for the validation set can be even more purposive if we are trying to extract maximum value from each sample, which is commonly done using some form of 'active learning'.[23]

The ground-truth dataset serves the crucial role of assessing the performance of the LLM on the task of interest. Classification output is most commonly assessed using the F1 score, which is the harmonic mean of recall and precision. Recall is the number of model 'hits' (true positives, e.g., cases about direct effect) divided by true positives and false negatives (cases the model incorrectly identified as not being about direct effect), while precision is the number of true positives divided by true

[22] Michal Ovádek, Philipp Schroeder, and Jan Zglinski. 'Where Law Meets Data: A Practical Guide to Expert Coding in Legal Research' (2024) 3(4) *European Law Open* 820–48.

[23] See, for example, Mitchell Bosley, Saki Kuzushima, Ted Enamorado, and Yuki Shiraito, 'Improving Probabilistic Models in Text Classification via Active Learning' (2022) arXiv preprint arXiv:2202.02629. Active learning is normally used to train supervised machine learning models more efficiently, but the same principles would apply if we wanted to leverage gold-standard data alongside LLM labels.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

positives and false positives (cases the model incorrectly identified as being about direct effect). Matthew's correlation coefficient (MCC) is another useful option, as it explicitly targets accuracy in imbalanced datasets.

There is no single F1 or MCC value at which we can generally say that the output of LLMs is accurate, but values above 0.9 (maximum 1) tend to be practically widely accepted. A key concern emerging out of literature on statistical inference in recent years is the fact that most researchers ignore measurement error as a source of uncertainty in downstream tasks.[24] Social scientists, including empirical legal scholars, are most likely to use LLMs to produce labels or quantities which feed into subsequent, typically regression, analysis. The problem is that if unaccounted for, measurement error will bias the downstream inferences.

In order to tackle this issue, Egami et al. draw on the doubly-robust machine learning literature[25] to derive what they call a design-based supervised learning estimator.[26] Their estimator leverages the fact that LLM users are producing gold-standard datasets to evaluate their performance anyway. The gold-standard data can, however, also be used to improve downstream statistical inference. In the common setting where outcomes are measured from text with error – for example, asking an LLM whether the CJEU agreed with the Commission in a decision – we can use the gold-standard labels to learn a model that will construct bias-corrected LLM outcomes conditional on document covariates. The subsequent regression of text-based outcomes on covariates of interest has the advantage of both asymptotic unbiasedness and better coverage than using LLM outputs outright in downstream analysis.

10.6 SYNTHETIC SAMPLES AND SIMULATIONS

One completely new avenue of research opened by LLMs is the creation of synthetic data and increasingly realistic simulations of real-world processes and outcomes. The extensive and wide-ranging knowledge of LLMs of the world can

[24] Laura Battaglia, Timothy Christensen, Stephen Hansen, and Szymon Sacher, 'Inference for Regression with Variables Generated from Unstructured Data' (2024) arXiv preprint arXiv:2402.15585; Amir Feder, Katherine A. Keith, and Emaad Manzoor et al., 'Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond' (2022) 10 *Transactions of the Association for Computational Linguistics* 1138; Naoki Egami, Christian J. Fong, Justin Grimmer, Margeret E. Roberts, and Brandon M. Stewart, 'How to Make Causal Inferences Using Texts' (2022) 8(42) *Science Advances* eabg2652; Dean Knox, Christopher Lucas, and Wendy K. Tam Cho, 'Testing Causal Theories with Learned Proxies' (2022) 25 *Annual Review of Political Science* 419.

[25] Victor Chernozhukov, Denis Chetverikov, and Mert Demirer et al., 'Double/Debiased Machine Learning for Treatment and Structural Parameters' (February 2018) 21(1) *Econometrics Journal* C1–C68.

[26] Egami (n 18).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

be leveraged to create synthetic observations in order to test hypotheses of interest. This line of research will be more speculative than using LLMs for classification, but as models improve and their legal capabilities in specific areas of law become validated, LLMs are likely to see increased uptake for evaluating alternative legal scenarios and predicting outcomes.[27]

In survey research, LLMs are used to create synthetic samples of respondents' answers to survey questions.[28] This allows researchers to evaluate a study's likely result without having to contact real respondents. In a similar fashion, legal scholars could create synthetic cases, party briefs, and judgments and study the effects of manipulating one or multiple variables which might be difficult to study in real-world conditions. For example, EU law researchers might ask what the impact of parties' legal briefs on CJEU decisions is. With access to data in this domain being limited, and given considerable heterogeneity in parties' propensity to submit observations to the Court, we could ask an LLM to write synthetic legal briefs based on each party's and its legal representation's characteristics, given some case facts. We could subsequently have the LLM write judgments based on the case facts and the synthetic briefs received from the parties. Varying the number and type of briefs received, we could estimate the impact of a party's synthetic brief on the ruling.

With the researcher being able to both control every variable and repeat the process again and again with the same or different parameters, LLMs can be powerful tools for the simulation of legal processes, provided there is fidelity between the synthetic data and its real-world counterparts. Similar to validating classification output, establishing that a synthetic legal brief of, for example, the French government in *Cassis de Dijon* is comparable to the real thing is going to be the most important and difficult aspect of LLM-based simulation studies. The training and validation of LLMs specialising in the production of EU legal advice in the case of legal briefs and CJEU rulings in the case of judgments will likely be necessary for such simulations to achieve their full potential.

10.7 CHOOSING AN LLM

Although GPT-4 is currently the most popular series of LLMs, there is a plethora of ever-expanding options to choose from as large companies and startups compete for a share of the market. The considerations guiding the choice of which LLM to use for researchers are slightly different from commercial applications, however.

[27] Guhong Chen, Liyang Fan, and Zihan Gong et al., 'AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents' (2024) arXiv preprint arXiv:2408.08089; Zhitao He, Pengfei Cao, and Chenhao Wang et al., 'SimuCourt: Building Judicial Decision-Making Agents with Real-World Judgement Documents' (2024) arXiv preprint arXiv:2403.02959.

[28] Lisa P. Argyle, Ethen C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate, 'Out of One, Many: Using Language Models to Simulate Human Samples' (2023) 31(3) *Political Analysis* 337; Shirley Mathur, Yajuan Si, and Jerome P. Reiter, 'Fully Synthetic Data for Complex Surveys' (2023) arXiv preprint arXiv:2309.09115.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

First of all, in empirical research we care about replicability and understanding the tools we are working with. Closed-source LLMs like GPT-4 do not enable users to peer into how the model was trained and the architecture it uses to generate output. In addition, the ability of future scholars to replicate work using closed-source LLMs controlled by their owners is questionable, because there is no guarantee that the company will still be serving the LLM in the same version when the replication is attempted. Given widespread consensus about the need for science to be open and transparent, closed-source LLMs leave much to be desired. Luckily, there are a number of powerful open-source models, such as Llama 3, developed by Meta. Such models come with complete documentation of training and data. Researchers need to ensure that their methods and findings are replicable and trustworthy. This openness allows for scrutiny, enabling researchers to identify and mitigate any biases or errors in the model's outputs. Transparency also fosters a collaborative environment where legal scholars can share methodologies, validate each other's work, and build upon existing research. Such openness is essential for maintaining the rigour and credibility of empirical legal studies, ultimately contributing to more robust and reliable legal scholarship. In addition, open-source LLMs make it possible to fine-tune models for specific domains and tasks, something that should naturally appeal to EU empirical legal researchers.

The closed or open nature of LLMs speaks to ethical concerns about their use as well. Providers who do not disclose the data and code used for their training obscure what may have amounted to questionable (privacy and copyright violating) data collection practices. Lack of access to the training data also prevents a deeper examination of the sources of bias in the output given by the LLM. Relatedly, using closed, proprietary LLMs means sharing potentially sensitive data with the provider company, the exploitation of which for further LLM training might require explicit opt-outs. In short, the training and deployment of LLMs, even in research, raises a plethora of ethical and legal questions.

Another consideration of particular importance to empirical legal scholars in the EU is LLMs' multilingual performance. Member States' jurisdictions bring EU law into direct contact with over twenty national languages, which traditionally created hurdles for comparative analyses of, for example, national implantation of EU directives. All LLMs are to some extent multilingual, but their performance across different languages varies as a function of their size (in parameters), architecture, and training data.[29] The performance drop-off is particularly noticeable for low-resource languages. While gap between the most commonly used European languages – English, French, German – and the rest is likely to persist, the vast majority of European languages have enough text data in them to train performant LLMs.

[29] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du, 'Quantifying Multilingual Performance of Large Language Models Across Languages' (2024) arXiv preprint arXiv:2404.11553.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013

One reason for this is the EU itself, as thousands of dense texts emanating from the EU's institutions – laws, speeches, and policy documents – are translated into smaller European languages.

10.8 CONCLUSION

The integration of LLMs into the empirical study of EU law represents a significant leap forward in the capabilities and methodologies available to legal scholars. These models, with their remarkable ability to process and generate human-like text, offer unprecedented efficiency in analysing vast corpora of legal documents, uncovering patterns, and providing data-driven insights. Despite their transformative potential, the use of LLMs in legal research is not without challenges. Concerns about the accuracy and reliability of LLM-generated outputs and the necessity for human validation are critical issues that must be addressed by every researcher using LLMs. There will also undoubtedly be technical barriers to adoption in the legal research community, in particular, as a result of its overall poorer training than other social science disciplines. Nonetheless, it is likely that at least some of these barriers will be overcome through the development of bespoke software applications – of which ChatGPT was perhaps the most successful example – which will require no knowledge of programming to use.[30]

Looking ahead, the future of LLMs in empirical legal studies hinges on the careful balance between leveraging their capabilities and mitigating their limitations. As the technology continues to evolve – we have not touched upon the multimodal nature of many LLMs, for example – it will be essential for legal scholars to develop methodologies for validating and complementing LLM outputs, ensuring that the insights derived are reliable. Moreover, the ongoing refinement of LLMs, coupled with the exploration of hybrid approaches such as RAG, promises to enhance their applicability and performance in legal research. By embracing these advancements while maintaining a critical perspective, the empirical study of EU law can harness the full potential of LLMs, driving forward our understanding of legal systems and decision-making.

[30] These are likely to raise transparency and reproducibility issues, however. There is no free lunch.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 16:04:38, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009672580.013