

## Natural Language Processing in Legal Tech

*Jens Frankenreiter and Julian Nyarko*

While the work of lawyers long appeared to be beyond the reach of automation, the “legal tech” revolution now seems to be in full swing. Of particular importance is the emergence of a new generation of legal tech applications that utilizes artificial intelligence (AI) and machine learning (ML). Their underlying technologies have started to profoundly change the work of various professionals, including bankers<sup>1</sup> and physicians.<sup>2</sup> And so it comes as no surprise that the rise of AI is also predicted to allow for the automation of some of the core tasks performed by lawyers.<sup>3</sup> Others go further, arguing that AI will ultimately allow computers to replace attorneys<sup>4</sup> and judges<sup>5</sup> in many scenarios, with profound changes for the functioning of the legal system.<sup>6</sup>

In reality, however, the extent to which AI can contribute to automation of the legal industry will depend on several factors, including regulation,<sup>7</sup>

<sup>1</sup> Jack Kelly, *Wells Fargo Predicts That Robots Will Steal 200,000 Banking Jobs within the Next 10 Years*, FORBES (Oct. 8, 2019, 11:09 AM EDT), <https://www.forbes.com/sites/jackkelly/2019/10/08/wells-fargo-predicts-that-robots-will-steal-200000-banking-jobs-within-the-next-10-years/?sh=5a9495db68d7>.

<sup>2</sup> David Killock, *AI Outperforms Radiologists in Mammographic Screening*, 17 NATURE REVIEWS. CLINICAL ONCOLOGY 134, 134 (2020).

<sup>3</sup> See, e.g., David Freeman Engstrom & Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. PA. L. REV. 1001, 1001–1004 (2021).

<sup>4</sup> E.g., John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Service*, 82 FORDHAM L. REV. 3041 (2014).

<sup>5</sup> E.g., Elliott Ash, *Robot Judges*, YOUTUBE (Feb. 13, 2018), <https://youtube.com/watch?v=6q1j7xSZKdo>.

<sup>6</sup> Benjamin Alarie, *The Path of the Law: Towards Legal Singularity*, 66 U. TORONTO L.J. 443, 443 (2016).

<sup>7</sup> See, e.g., Gillian K. Hadfield, *Legal Barriers to Innovation: The Growing Economic Cost of Professional Control over Corporate Legal Markets*, 60 STAN. L. REV. 1689, 1724–25 (2008).

culture,<sup>8</sup> and technology.<sup>9</sup> In this chapter, we focus on the last factor, the current state and likely future trajectory of technological progress. Among the different technological requirements, our focus is Natural Language Processing (NLP), a key component of many current and envisioned legal tech tools. Understanding the inherent constraints of current NLP models will be crucial in determining the extent to which legal tech applications will succeed in their quest to revolutionize the market for legal services. Although it cannot be ruled out that “robot-judges” and “robo-lawyers” will one day dominate the reality of legal engagement, our analysis suggests that recent developments in the relevant fields provide no basis for the prediction that such comprehensive legal automation is right around the corner.

The legal system trades in words, and NLP promises to automate an activity that lies at the heart of many tasks performed by lawyers: the extraction and processing of information from unstructured text. Lawyers routinely encounter unstructured text in their daily work routine, be it in the form of judicial opinions, statutes, legal briefs, written agreements, or witness testimony. Understanding and processing the information from this text is essential for them to be effective. For example, without reading prior case law, lawyers will generally be unable to determine whether a case at hand has a chance of succeeding in court. Consequently, many legal tech applications, and particularly those seeking to automate the tasks lying at the heart of what it means to “be a lawyer,” depend on NLP to process such information in a meaningful way.

We provide readers with an overview of the current state of NLP techniques, focusing on their promise and potential pitfalls in the context of legal tech applications. Like many other fields of AI, NLP has seen some drastic improvements in recent decades. Among other things, these improvements have contributed to some of the most talked-about success stories in legal tech, including the development of novel tools that promise to facilitate often mundane and repetitive tasks like document review. Against this background, one might conclude that advances in NLP are poised to similarly benefit the development of other legal tech tools. Yet, while there certainly is ample room for optimism, a realistic outlook must simultaneously recognize that NLP suffers inherent and important constraints, limiting its utility for legal tech applications in significant ways.

We explain why NLP-powered legal tech can be expected to outperform humans in specific categories of tasks that play to the strengths of current ML techniques. In particular, NLP-assisted applications perform well in prediction or classification

<sup>8</sup> John Armour et al., *Augmented Lawyering* (Eur. Corp. Governance Inst. Working Paper, Paper No. 558, 2020), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3688896](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3688896).

<sup>9</sup> See Milan Markovic, *Rise of the Robot Lawyers?* 61 ARIZ. L. REV. 325, 328–35 (2019).

tasks in which large amounts of pre-labeled data is available or can be generated to train an algorithm. The availability of suitable training data enables an algorithm to detect even subtle patterns in text or other data that predict the label a human would have attached to the document or other item.<sup>10</sup> At the same time, the chapter explores and highlights some of the central obstacles to deploying these tools in other contexts, most importantly in tasks that require the equivalent of legal reasoning. This includes, for instance, applications seeking to assess the likely outcome of legal disputes.

To be successful in this latter context, legal tech applications would need to automatically extract from relevant texts a structured representation of legal concepts and their interconnections (what we refer to as a “legal ontology”). However, despite recent progress in developing tools that appear – at least on a superficial level – capable of extracting meaning and knowledge from text, attempts to derive legal ontologies have so far been largely unsuccessful. On the contrary, recent studies suggest that current advancements in NLP, for the most part, do not meaningfully increase the performance of algorithms in tasks requiring legal reasoning. This, in turn, suggests that many legal tech applications may not benefit significantly from improvements made in general language processing and general language understanding. Instead, in order to make significant progress, a concentrated and domain-specific effort may be required that is specifically designed to promote the capabilities of language models to engage in forms of legal reasoning.

### 3.1 NLP AS PART OF THE BROADER LEGAL TECH LANDSCAPE

#### 3.1.1 *The Rise of ML-Powered Legal Tech*

Changes to the work of lawyers brought about by technology are not a new phenomenon. Over the last decades, the introduction of technologies such as email and word processing and the availability of computer-based legal databases have profoundly changed lawyers’ daily routines. With the advent of ever-more-powerful computers, the Internet, and recent developments in AI and ML, this development appears to have accelerated. The emerging field of “legal tech” promises to equip lawyers with tools capable of automatically generating substantive content descriptions of contract clauses,<sup>11</sup> analyzing legal briefs,<sup>12</sup> and performing data-driven venue

<sup>10</sup> See, e.g., AJAY AGRAWAL ET AL., *PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE* (2018).

<sup>11</sup> *How Kira Works*, KIRA, <https://kirasystems.com/how-kira-works/>.

<sup>12</sup> *Legal Analytics*, BLOOMBERG L., <https://pro.bloomberglaw.com/ai-analytics/>.

analysis.<sup>13</sup> In addition, because of its potential to scale, technology is often seen as a way to make legal services available to those who cannot afford a (human) lawyer.<sup>14</sup>

Many commentators writing about legal tech adopt a broad definition that goes well beyond applications seeking to automate the activities at the core of a lawyer's work. As an example, consider a recent article by David Engstrom and Jonah Gelbach. Their definition of legal tech encompasses applications ranging from outcome prediction to online marketplaces for lawyers.<sup>15</sup> Notably, as the example of online marketplaces for lawyers illustrates, not all of these applications relate to tasks that lawyers routinely perform as part of their work. And even within the set of applications that substitute for work usually performed by lawyers, the technical sophistication required to automate the task varies substantially.<sup>16</sup>

Despite this broad definition of legal tech, many observers are particularly interested in legal tech applications made possible by applying ML and similar techniques to the legal field.<sup>17</sup> While legal tech includes applications powered by technology that has been around for decades (in one form or another), many tasks performed by lawyers and other knowledge workers long seemed beyond the reach of automation. Recent advances in ML, however, have enabled significant technological progress in many areas that were commonly considered the sole domain of humans, including the driving of cars,<sup>18</sup> translation,<sup>19</sup> and the writing of human-readable text<sup>20</sup> and music.<sup>21</sup> Not surprisingly, these developments have also spurred discussion about novel legal tech applications that automate the work of lawyers to a hitherto unknown extent.<sup>22</sup>

<sup>13</sup> *Legal Analytics Platform*, LEX MACHINA, <https://lexmachina.com/legal-analytics/>.

<sup>14</sup> Zorik Pesochinsky, *Leveraging Legal Technology to Improve Access to Justice*, THOMSON REUTERS (July 29, 2019), <https://www.thomsonreuters.com/en-us/posts/legal/leveraging-legal-tech-access-to-justice/>.

<sup>15</sup> Engstrom & Gelbach, *Legal Tech*, at 1011–12.

<sup>16</sup> *Id.* at 1014.

<sup>17</sup> *E.g.*, *id.* at 1015 (“Nor should it surprise that the most promising legal tech tools deploy ML.”).

<sup>18</sup> Joe D’Allegro, *How Google’s Self-Driving Car Will Change Everything*, INVESTOPEDIA (June 20, 2021), <https://www.investopedia.com/articles/investing/052014/how-googles-selfdriving-car-will-change-everything.asp>. Recently, it has become clear that some earlier predictions about the imminent arrival of self-driving cars were overoptimistic. Neal E. Boudette, *Despite High Hopes, Self-Driving Cars Are “Way in the Future.”* N.Y. TIMES (July 17, 2019), <https://www.nytimes.com/2019/07/17/business/self-driving-autonomous-cars.html>.

<sup>19</sup> Gideon Lewis-Kraus, *The Great A.I. Awakening*, N.Y. TIMES MAG. (Dec. 14, 2016), <https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html>.

<sup>20</sup> Will Knight, *This AI Can Generate Convincing Text – and Anyone Can Use It*, WIRED (Mar. 29, 2021 7:00 AM), [www.wired.com/story/ai-generate-convincing-text-anyone-use-it/](http://www.wired.com/story/ai-generate-convincing-text-anyone-use-it/).

<sup>21</sup> Parker Hall, *Dynascore’s AI Music Engine Writes Tracks to Match Your Videos*, WIRED (July 7, 2021, 7:00 AM), <https://www.wired.com/story/dynascore-ai-music-engine/>.

<sup>22</sup> Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics*, 68 U. TORONTO L.J. 12, 27 (2018); Daniel Martin Katz, *Quantitative Legal Prediction – Or – How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909 (2013).

Many techniques in ML are designed to generate predictions from example data.<sup>23</sup> Consequently, its potential is greatest for tasks that can comfortably be viewed as prediction exercises. To illustrate, consider reviewing a large number of documents with the aim to identify privileged information before the documents are shared with the opposing party in litigation.<sup>24</sup> This task can comfortably be characterized as a prediction exercise: The goal is to determine whether a document should receive one of two labels, “privileged” or “non-privileged.” ML techniques can facilitate this process. In particular, through supervised classification, an ML algorithm can be trained to “learn” the difference between privileged and non-privileged documents using a small, human-labeled sample. Once trained, the algorithm can then create predictions for the entire, unlabeled corpus of documents. In this way, the ML algorithm “predicts” the label that is most likely to be assigned to a document by a human coder, based on the features of the labeled documents.<sup>25</sup>

Another task that is often discussed as an important use case for ML-assisted legal tech is outcome prediction. Potential litigants deciding whether to file a case – and attorneys deciding whether to represent them – will usually attempt to form an expectation about the chances of succeeding at trial. At least in principle, this is a straightforward prediction task: If the right data were available, ML algorithms should be able to determine whether a plaintiff is likely to win by comparing the features of the present case with the features of past cases.<sup>26</sup>

Of course, the examples above do not constitute an exhaustive list of potential applications. Instead, they simply serve as an illustration for the broad range of tasks that may benefit from developments in NLP.

<sup>23</sup> Marion Dumas & Jens Frankenreiter, *Text as Observational Data*, in LAW AS DATA 59 (Michael A. Livermore & Daniel N. Rockmore eds., 2019); Jon Kleinberg et al., *Prediction Policy Problems*, 105 AM. ECON. REV. 491 (2015); Sendhil Mullainathan & Jann Spiess, *Machine Learning: An Applied Econometric Approach*, J. ECON. PERSPS. (Spring 2017), at 87.

<sup>24</sup> See Chapters 5 and 6 in this volume.

<sup>25</sup> See also Dana A. Remus, *The Uncertain Promise of Predictive Coding*, 99 IOWA L. REV. 1691, 1701–706 (2014).

<sup>26</sup> Benjamin Alarie et al., Using Machine Learning to Predict Outcomes in Tax Law (Dec. 15, 2017) (unpublished manuscript), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2855977](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2855977); Nicole Black, Finding Treasure with Litigation Data Analytics Software, A.B.A. J. (July 24, 2018, 6:05 AM CDT), [https://www.abajournal.com/news/article/finding\\_treasure\\_with\\_litigation\\_data\\_analytics\\_software](https://www.abajournal.com/news/article/finding_treasure_with_litigation_data_analytics_software); Katz, *Quantitative Legal Prediction*, at 936–42; Dru Stevenson & Nicholas J. Wagner, *Bargaining in the Shadow of Big Data*, 67 FLA. L. REV. 1337, 1371–74 (2016). There is also a considerable literature at the intersection of law and data science that aims to develop and test outcome prediction models in the context of various courts. See, e.g., Daniel Martin Katz et al., *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, 12 PLoS ONE 1 (2017); Masha Medvedeva et al., *Using Machine Learning to Predict Decisions of the European Court of Human Rights*, 28 A.I. & L. 237 (2020).

### 3.1.2 The Role of NLP in Legal Tech

NLP, as used here, refers to a set of methods that allow computers to process and extract information from human language. NLP constitutes an important building block for various programming tasks, including programs enabling computers to act on commands issued in spoken language<sup>27</sup> and translate text between different languages.<sup>28</sup> Like much other AI research, NLP has been a topic of inquiry (in one form or another) since at least the 1950s. In earlier decades, and in line with the then-governing paradigm guiding much research in AI, NLP operated on the basis of a “top down” approach: Computers extracted information from texts following complex sets of hand-coded rules drawn up by linguists and other experts.<sup>29</sup> Since the 1990s, NLP has undergone a profound transition that mirrors similar developments in other areas of AI. Today, most NLP follows a “bottom up” approach, relying heavily on ML and similar statistical techniques. These approaches allow computers to “learn” how to process language on the basis of large amounts of training data.<sup>30</sup>

On a very general level, NLP techniques hold the potential to automate an activity that plays a central role in many tasks performed by lawyers: the extraction and processing of information from unstructured text, either spoken or written. By unstructured text, we mean any text that is not organized in a manner that readily maps onto predefined categories known to be relevant for the task at hand. Most information encountered by lawyers in their daily work takes the form of unstructured text. Examples include the text of judicial opinions, statutes, legal briefs or written agreements, or written witness testimony. In contrast, the information provided by a client in response to an intake questionnaire that allows for only a predefined set of responses to each question does not constitute unstructured text, and NLP is usually not needed to process this kind of information.

To illustrate the central role of information extraction from unstructured text, consider again the example of outcome prediction. An attorney trying to predict whether a case presented to her by a client will succeed in court will usually require at least two types of information: first, information about the facts of the case and, second, information about how courts treat cases like the one at hand. The latter category includes, in particular, information about the applicable legal rules.<sup>31</sup> Both

<sup>27</sup> Bill Manaris, *Natural Language Processing: A Human-Computer Interaction Perspective*, 47 *ADVANCES COMPUTS.* 1 (1998).

<sup>28</sup> Lewis-Kraus, *The Great A.I. Awakening*.

<sup>29</sup> See, e.g., Antoine Louis, *A Brief History of Natural Language Processing – Part 1*, *MEDIUM* (July 7, 2020), <https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbc937ebce>.

<sup>30</sup> *Id.*

<sup>31</sup> Other factors that might influence how a court will likely treat cases like the one at hand include information about judges, juries, and political trends – in other words, “external” factors. See also Engstrom & Gelbach, *Legal Tech*, at 1027.

types of information will primarily be available in the form of unstructured text. Information about the facts of the case often comes in the form of written or spoken statements by the client, potentially combined with additional documents such as contracts; it can also involve depositions of witnesses and background research. Information about the law will typically take the form of statutes, regulations, and other cases, plus (depending on the jurisdiction) secondary sources of law such as legal treatises. Only after obtaining information about both the facts and the law (and potentially other factors that might influence the decision) will the attorney be able to determine the chances to succeed in trial.

This example is not an outlier. Instead, the processing of unstructured text plays a central role in many tasks performed by lawyers. A lawyer conducting document review reads documents to determine whether they fall into certain predefined categories (relevant to the case at hand, privileged, etc.) A lawyer drafting an agreement not only researches the law to establish the legal framework against which the contract is set, but also consults templates or examples of similar agreements she or her colleagues have drafted in the past.

Because of the central role that the extraction and processing of information from unstructured text plays in the work of lawyers, NLP is a potential key component of many actual or hypothesized legal tech applications.<sup>32</sup>

NLP will likely be particularly important in applications that seek to increase access to legal services to those who cannot afford a lawyer. In all the examples described above, before even starting their work on a case, lawyers need to decode unstructured language (often in the form of spoken text) to understand their clients' goals. Insofar as legal tech applications seek to interact directly with clients on the basis of unstructured language descriptions provided by the latter, NLP will also be needed to automate the task of translating these narratives into legally relevant units of information.<sup>33</sup> In addition, NLP can play an important role in communicating the results of the work performed by a legal tech application to a client in language that she can understand.

<sup>32</sup> See also *id.* at 1013; Jens Frankenreiter & Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANN. REV. L. SOC. SCI. 39, 51 (2020); Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics*, 68 U. TORONTO L.J. 12, 27 (2018).

<sup>33</sup> We note that the task of direct client interaction through unstructured input text is particularly challenging and requires a combination of intent extraction (i.e., the system's ability to extract what a user wants when making a query), Named Entity Recognition, and language parsing. See also Chapter 2 in this volume. An alternative (and often more feasible) solution can be to limit the input provided by the user to a narrow set of predefined options. This approach is particularly popular in narrow domains with clear rules, such as tax software. Instead of relying on open-ended text, these applications help clients to compute their tax burden by iteratively querying relevant information through a series of questions that allow for a limited set of possible answers.

TABLE 3.1 *Sample vectors in a BoW model*

| Word      | Sentence 1 | Sentence 2 | Sentence 3 |
|-----------|------------|------------|------------|
| we        | 1          | 1          | 1          |
| protect   | 1          | 0          | 0          |
| safeguard | 0          | 1          | 0          |
| expose    | 0          | 0          | 1          |
| your      | 1          | 1          | 1          |
| data      | 1          | 1          | 1          |

### 3.2. THE NLP PREDICTION PIPELINE

Typically, the process of creating predictions from text consists of two steps. In the first step, a language model converts text into vectors. In a second step, these vectors are used as an input into a ML classifier in order to create relevant predictions.<sup>34</sup> We discuss each step in turn.

#### 3.2.1 *Step 1: Converting Text into Vectors Using Language Models*

Available language models differ in terms of their complexity. At the lower end of the spectrum are simple language models such as the “Bag-of-Words” (BoW) model.<sup>35</sup> These models do not involve a formal training process. Instead, each word is assigned its own, unique vector representation. Take, for instance, the following three sentences:

*Sentence 1: We protect your data*

*Sentence 2: We safeguard your data*

*Sentence 3: We expose your data*

To represent these sentences, a BoW model creates three vectors, one for each sentence. The vectors have as many dimensions as there are unique words (here 6). The elements of the vectors indicate whether a word is present or not. In this case, the column vectors for the three sentences are given in Table 3.1.

The above representation illustrates two important aspects of the BoW model: First, there is a lot of overlap between the vector representations of the sentences. Each vector has a 1 in the first, fifth, and sixth position. Hence, this simple language model accurately suggests that all three sentences are somewhat related, in the sense that they all talk about how another party handles the addressee’s data. At the same time, however, the BoW representation has a significant shortcoming. In particular, the overlap between Sentence 1 and Sentence 2 is the same as the overlap between

<sup>34</sup> In modern language models, both steps may interact somewhat with each other.

<sup>35</sup> Yin Zhang et al., *Understanding Bag-of-Words Model: A Statistical Framework*, 1 INT’L. J. MACH. LEARNING & CYBERNETICS 43, 43 (2010).

Sentence 1 and Sentence 3. However, any human reader would attest that the words *protect* and *safeguard* are semantically close, whereas *protect* and *expose* have almost polar opposite meanings. In this sense, a simple BoW model may be able to capture some semantic similarity between texts. However, since it simply represents word occurrence, it does not encode information about the semantic meaning of text. If we were to use the output of this BoW in a downstream task, it could quickly lead to incorrect results, particularly for nuanced tasks in which every single word carries significance.

To address this problem, current language models seek to identify numerical representations that are more faithful to the semantic meaning of a word or sentence. Under these models, Sentence 1 and Sentence 2 would be represented by vectors with similar elements, whereas Sentence 1 and Sentence 3 are represented in a way that is faithful to the difference in their semantic meaning. The details on how these models are able to achieve their goal differ, but what all modern language models have in common is a reliance on distributional semantics. The intuition behind distributional semantics is that linguistic items with similar semantic meanings are also distributed similarly. Indeed, John Firth famously said that “a word is characterized by the company it keeps,”<sup>36</sup> a linguistic notion similarly supported by Wittgenstein, who argued that the meaning of a word can only be understood by learning how said word is used in context.<sup>37</sup> We can easily see that this assumption is often reasonable. Assume, for instance, that we encounter a new word spelled *quari*. Without context, it would be difficult to tell what this word means. But now assume we encounter the word being used in practice. Perhaps we encounter sentences such as “I took my *quari* to the mechanic yesterday,” or “I was driving fast with my *quari* on the highway.” At this point, we could start forming expectations that a *quari* is a form of vehicle, because it is used in a way that is similar to how we use words such as *car* or *automobile*. At some point, we may even start using *quari* as a synonym for *car*.

The objective on which modern language models are trained imitates this type of learning from distributional properties. For instance, Google’s BERT and OpenAI’s GPT-3 – two recent large language models that have been hailed as breakthrough innovations for NLP – are trained simply by repeatedly guessing which word is most likely to occur next in a sequence of words.<sup>38</sup> The idea behind this training objective is straightforward: If a model is optimized to predict the next word in a sentence, it

<sup>36</sup> J. R. Firth, *A Synopsis of Linguistic Theory, 1930–1955*, in *STUD. IN LINGUISTIC ANALYSIS* 1 (1957).

<sup>37</sup> LUDWIG WITTGENSTEIN, *PHILOSOPHICAL INVESTIGATIONS* (1953).

<sup>38</sup> Jacob Devlin et al., *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*, ARXIV (2019), <https://arxiv.org/abs/1810.04805>; Tom B. Brown et al., *Language Models Are Few-Shot Learners*, 33 *ADVANCES NEURAL INFO. PROCESSING SYS.*, 2020, at 1. In addition, Google’s BERT guesses which of two candidate sentences follows the input.

inherently learns linguistic properties that correspond to semantic meaning. For instance, a model learning that it is likely that the next word in the sentence “I was driving fast with my . . . ” is *car* or *motorcycle* but not *cat* implicitly learns that a *motorcycle* is closer in meaning to *car* than it is to *cat*. And it turns out that next word prediction at scale is able to implicitly encode much more than just semantic meaning. For instance, by repeatedly solving next word prediction tasks, language models can “learn” grammatical rules<sup>39</sup> and encode world and limited legal knowledge.<sup>40</sup>

### 3.2.2 Step 2: Downstream Prediction Tasks

Once a language model has been trained, it allows for the conversion of text into vectors that encode the semantic meaning of a word, sentence, or document. These vectors can then be used in a ML model tasked with generating predictions that can vary with the objective of the legal tech application.

Like all ML models, those used in the context of NLP applications require at least two different types of data: training data and input data.<sup>41</sup> Training data refers to the information that is used to “calibrate” the statistical models forming the core of the ML tools. In other words, the training data allows the algorithm to “learn” about the relationship between some input and the desired prediction. For many language-based legal tech applications, training data consists of texts alongside labels that reflect certain information contained in the text. During training, the algorithm will learn which textual cues are strongly associated with the individual labels.<sup>42</sup> For example, a legal tech application assisting in electronic discovery may seek to automatically distinguish between documents that do and do not contain privileged information. In order to achieve that goal, the creator of the legal tech application would begin by hand-labeling a training corpus of documents for whether they contain privileged information or not. She would then use a language model to transform the content of the training documents into numerical vectors and would feed these vectors into a ML algorithm that learns the relationship between the different elements of the vectors and the human annotation (privileged/non-privileged).

Input data is the data that is fed into the machine to generate the predictions or other results that the user of the legal tech application is interested in. Notably, and

<sup>39</sup> A model that correctly predicts that the word *drove* is more likely to occur than the word *drive* in a sentence that begins with “Yesterday, I” implicitly encodes that *drove* is the past tense of *drive*.

<sup>40</sup> A model knowing that “The plaintiff sued the” will be followed by “defendant” implicitly learns that most cases have both a plaintiff and a defendant.

<sup>41</sup> AGRAWAL, PREDICTION MACHINES, at 43.

<sup>42</sup> In principle, labels can consist of unstructured text as well (so-called sequence-to-sequence labeling).

in contrast to the training data, input data can be largely unlabeled. In the context of document review, the input data usually consists of the unlabeled documents that were not included in the training dataset. After the ML algorithm has been trained, it can generate predictions for these documents at scale.

### 3.2.3 Two Use Cases

To illustrate both the utility and limitation of NLP in legal tech, we focus on two applications – document review and case outcome prediction. Both tasks lie at opposite ends of the spectrum of legal cognitions and thus bring some of the problems of NLP in sharp relief. At the same time, we are conscious of the fact that the full spectrum of legal tech implementations encompasses a much larger set of applications and that implications we draw here might apply differently in different use cases.

In the case of document review, the standard NLP pipeline described above performs well.<sup>43</sup> This is in part due to the fact that document review closely resembles information extraction, a standard linguistic prediction task that these models are particularly designed for. For instance, a typical exercise in NLP research is to predict the sentiment of social media posts on the basis of the posts' texts and a labeled training dataset.<sup>44</sup> Similarly, as described above, document review can be understood as an exercise in predicting labels associated with a document (e.g., privileged/non-privileged) on the basis of the text of the document.

Other tasks, however, are much more difficult to perform, irrespective of how sophisticated the language model is. One such task is the prediction of the outcome of a legal dispute. Legal outcome prediction using expert-generated systems has a long tradition.<sup>45</sup> However, here, we focus on the automatic prediction of legal outcomes without any human intervention.<sup>46</sup>

To illustrate the difficulties associated with legal outcome prediction, consider how the standard NLP pipeline might be implemented for this task.<sup>47</sup> At first, the process may appear to be straightforward: A language model could transform textual descriptions of fact patterns into vectors, with each vector in the training dataset

<sup>43</sup> See Engstrom & Gelbach, *Legal Tech*, at 1021; see also Chapter 5 in this volume.

<sup>44</sup> Lin Yue et al., *A Survey of Sentiment Analysis in Social Media*, 60 *KNOWLEDGE & INFO. SYS.* 617, 617–63 (2019).

<sup>45</sup> For one historical account, see Kevin D. Ashley, *A Brief History of the Changing Roles of Case Prediction in AI and Law*, *L. CONTEXT FOR DIGIT. AGE*, 2019, at 93, 93–112.

<sup>46</sup> For recent studies in legal outcome prediction, see Medvedeva, *Using Machine Learning*, at 237–66; Haoxi Zhong et al., *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence*, *ARXIV* (2020), <https://arxiv.org/abs/2004.12158>; L. Karl Branting et al., *Scalable and Explainable Legal Prediction*, 29 *A.I. L.* 213, 213–38 (2021); Elizabeth C. Tippet et al., *Does Lawyering Matter? Predicting Judicial Decisions from Legal Briefs, and What That Means for Access to Justice*, 100 *TEXAS L. REV.* (forthcoming 2022).

<sup>47</sup> See Chapter 7 in this volume.

accompanied by a label indicating whether the plaintiff or the defendant won the case.<sup>48</sup> The ML algorithm would then generate predictions for the input dataset depending on the textual features of the document. The approach may sound simple in theory. And yet, to date, NLP-based attempts at predicting the outcome of legal cases using the standardized process have failed to produce reliable results.<sup>49</sup> But what is it that differentiates legal outcome prediction tasks from other exercises, such as classifying documents into relevant and non-relevant?

At its core, the difference lies in the quantity of linguistic markers that influence the prediction task. In legal outcome prediction, the set of informative linguistic markers is potentially infinite, whereas in document review, there typically is a small set of linguistic cues the algorithm is required to pick up on. To illustrate, imagine that there was a universal speed limit of fifty miles per hour. If we trained a model to predict the outcome of a dispute over a citation, it would quickly pick up on the fact that the most important linguistic cue for the prediction of the outcome is the description of the car's speed. If that description contains phrases such as "sixty mph," it would predict that a legal challenge to the citation is unlikely to succeed, whereas phrases such as "thirty mph" would lead the algorithm to predict a higher success rate. Perhaps the model would alter its prediction based on whether the description indicates that a radar was used or whether the speed was measured through pacing. But overall, the possible combinations of types of fact patterns that are relevant to the prediction are finite and small, and it is likely that (almost) every type of fact pattern relevant to the prediction could be sufficiently represented in a large enough corpus. Importantly, note that the classifier could achieve a high performance without ever being told what the specific rule is. Indeed, given a large enough training sample, the algorithm would automatically identify that "fifty mph" appears to be the discontinuity at which the probability of success changes significantly.

However, most legal prediction tasks, and certainly most of the law applicable to disputes, are not as simple. Consider, for instance, a negligence tort dispute. In contrast to the clear fifty mph rule in the narrow context of driving on public roads, negligence is a vague standard that may trigger liability in an unquantifiable number of contexts. The indefinite number of variations in fact patterns comes with a similarly indefinite number of potential linguistic cues that are relevant to the prediction of the outcome. When designing a legal technology app, it is impossible

<sup>48</sup> We ignore for now difficulties in defining who "won" a case, which certainly is a context-specific analysis.

<sup>49</sup> For an overview, see Zhong, *How Does NLP Benefit Legal System*. To be sure, in some applications, researchers report F1 scores in the range of 0.7–0.8 when predicting the outcome of judicial opinions. See, e.g., Medvedeva, *Using Machine Learning*, at 237–66; Tippett, *Does Lawyering Matter?* However, these studies are based on published opinions and/or use as input data documents that are produced at later stages in the litigation process. As such, they are best considered "backward looking," providing ex post insights on how cases have been decided, rather than on how future cases will be decided.

to collect a training dataset that contains all potential fact patterns that may or may not lead to liability under a negligence regime. In this scenario, a language model does not have the opportunity to detect and assess the relevance of all these linguistic cues that might be relevant to the success rate of the plaintiff. In order to make accurate predictions, the algorithm can no longer rely on comparing the linguistic and distributional patterns of one document to other documents it has encountered during training. Rather, in order to be successful, the algorithm has to perform the equivalent of *legal reasoning*. In other words, to fully automate the process without significant human intervention, an algorithm would have to be able to extract from the texts a representation of legal concepts and their interconnections (a “legal ontology”). With such a legal ontology, the algorithm would then be able to draw accurate inferences from precedent, even if it had never encountered the particular fact pattern in the past. For instance, if an algorithm was able to infer that fact patterns describing a small burden and a large probability of a significant loss increase the chances for the plaintiff to prevail in a negligence suit, it could use this knowledge to inform its predictions without reliance on the particular words used in the description of facts.

Against this background, one might ask whether the impressive trajectory of modern language models in recent years also translates into an ability to reconstruct such legal ontologies from text.<sup>50</sup> Unfortunately, the results to date are not promising. In fact, existing algorithms are notoriously bad at extracting abstract rules from text, and legal ontologies are no exception. Even if a researcher feeds the text of the rules directly into the model, current language models are not able to accurately process that information. For instance, Holzenberger and colleagues<sup>51</sup> task a language model with predicting how much tax is owed by an individual, using as input a description of facts. They show that, even though the relevant ontology can be defined manually by an expert and used to compute the exact tax rate, language models perform poorly at this task. But perhaps more importantly, the authors point out that model performance does not increase even after supplementing the description of facts in the training data with the text of the relevant statutes. This suggests that language models have difficulties deriving a legal ontology from text without human assistance.

To corroborate these intuitions, we turned to GPT-3, a state-of-the-art language model that has achieved extraordinary results on a number of language tasks.<sup>52</sup> Among others, it has demonstrated its ability to translate a description of a computer

<sup>50</sup> Bonan Min et al., *Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey*, ARXIV (2021), <https://arxiv.org/abs/2111.01243>; Zhong, *How Does NLP Benefit Legal System*, at 5.

<sup>51</sup> Nils Holzenberger, Andrew Blair-Stanek & Benjamin Van Durme, *A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering*, ARXIV (2020), <https://arxiv.org/abs/2005.05257>.

<sup>52</sup> Brown, *Language Models*.

program from ordinary language into a working software application,<sup>53</sup> and it has written blog posts that fooled thousands of people into thinking they were authored by a human.<sup>54</sup> The legal literature has begun to study the utility of GPT-3 for tasks such as contract analysis in the consumer context, with at times positive results.<sup>55</sup>

To assess the capabilities of GPT-3 for a simple task involving legal reasoning, we first submitted the following query:<sup>56</sup>

*Are liquidated damages clauses enforceable?*

The (correct) response:

*Liquidated damages clauses are generally enforceable unless the sum stipulated is exorbitant or unconscionable.*

However, we then submitted the following query:

*X and Y contract over the sale of a Toyota Corolla. Their contract specifies that X has to pay Y one million dollars in damages if X does not deliver the car on time. X does not deliver the car on time. Q: Is Y entitled to damages? If so, in what amount?*

GPT-3's response:

*Yes, Y is entitled to one million dollars in damages. Damages are meant to compensate the injured party for the loss suffered due to the breach. Y has suffered a loss of one million dollars because X has not delivered the car on time.*

GPT-3 has been trained on millions of documents including almost half a trillion words. And while it was able to correctly extract and restate the legal rule based on the documents it encountered during training, it was not able to convert this rule into a legal ontology that would enable it to correctly infer the invalidity of the contractual penalty.

Next, we explicitly specified that the stipulated damages were “exorbitant.” We also fed the relevant rule that GPT-3 itself recovered back into the model, resulting in the following query:

*X and Y contract over the sale of a Toyota Corolla. Their contract specifies that X has to pay Y the exorbitant amount of one million dollars in damages if X does not deliver the car on time. X does not deliver the car on time. The law requires that liquidated damages clauses are enforced unless the sum stipulated is exorbitant or unconscionable. Q: Is Y entitled to damages? If so, in what amount?*

<sup>53</sup> @jsngr, TWITTER (July 18, 2020, 8:31 AM), <https://twitter.com/jsngr/status/1284511080715362304>.

<sup>54</sup> Liam Porr, *My GPT-3 Blog Got 26 Thousand Visitors in 2 Weeks*, EXCAVATIONS (Aug. 3, 2020), <https://liamp.substack.com/p/my-gpt-3-blog-got-26-thousand-visitors>.

<sup>55</sup> Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart Readers*, 90 GEO. WASH. L. REV. (forthcoming 2022).

<sup>56</sup> This example is a reprint one of us has used before to highlight the limitations of GPT-3 in Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ARXIV (2021), <https://arxiv.org/abs/2108.07258>.

GPT-3's response:

*Yes, Y is entitled to the liquidated damages. The amount in this case is not exorbitant since the price was set by the parties themselves.*

Again, the model ignores the legal rules and produces an incorrect prediction. To be sure, we emphasize that our queries are not meant to scientifically explore the limitations of GPT-3. Instead, our application is meant to exemplify a more general point. As we pointed out above, the training objective of language models causes them to encode information that is included in the distributional properties of written language, such as word and sentence order. Language models are thus able to excel at tasks that require accessing distributional information. For instance, in order to write a text that is indistinguishable from a text written by a five-year-old human, most of the relevant information is contained in the grammatical structure and frequency distribution of words.<sup>57</sup> However, language understanding in general, and legal reasoning in particular, requires processing of information that is not a mere reflection of linguistic patterns. Instead, lawyering requires a “dizzying array of analytic moves,”<sup>58</sup> including an ability to apply rules to facts, inferring regularities from existing case law, distinguishing new fact patterns from precedent, and using logical reasoning and creativity to craft new legal arguments. Modern language models do not even begin to achieve performance that is close to that of human professionals. Perhaps even more importantly, the performance of these language models does not appear to significantly increase with the sophistication or complexity of the language model.<sup>59</sup> This suggests that the current trajectory of developments in NLP, which relies heavily on distributional properties, may run orthogonal to the kinds of models that would be needed to create accurate predictions of legal disputes.

Together, the current developments point toward a more limited role for NLP in applications involving legal reasoning: It appears that the task of building, dissecting, and understanding the law in order to assess a novel set of facts will continue to rely primarily on domain experts.<sup>60</sup> Creators of legal tech applications will have to rely on these experts to define a knowledge system that determines the relevant factors under the law, as well as how these factors influence the likely outcome. This does not mean that NLP will be irrelevant for such applications. It could still be used in more narrowly defined information extraction tasks with the aim of determining whether the factors that matter for the outcome of a dispute are present in a case at hand.

<sup>57</sup> For instance, five-year-olds have a limited vocabulary and build simple sentences.

<sup>58</sup> Engstrom & Gelbach, *Legal Tech*, at 1024.

<sup>59</sup> Holzenberger, *A Dataset for Statutory Reasoning*.

<sup>60</sup> Engstrom & Gelbach, *Legal Tech*, at 1025–26.

### 3.3. OTHER KEY CHALLENGES

The challenge of automating the creation of legal ontologies makes it difficult to predict whether, and when, technology will be able to completely replace the prominent and essential role humans – and, in particular, human lawyers – play in creating a functioning legal regime. Of course, it is famously difficult to predict the capabilities of future AI systems in many different contexts. However, one thing we do know is that recent advancements in NLP, although in many ways impressive, did not move us significantly closer in automating many of the core tasks that lawyers perform on a daily basis. It may well be the case that the full automation of most legal services is not only years, but decades away. But this is not all. At least four further challenges will impose hurdles on the development of legal tech tools and shape the field's future trajectory.

#### 3.4.1 Document Structure and Segmentation

Before working with a text corpus, it needs to be broken up into coherent, informative units of analysis. This process is also referred to as “document segmentation” and is an important step in the NLP pipeline that can have significant consequences both for training a language model as well as for other downstream prediction tasks. In principle, the investigator is free to break text up at varying levels of granularity, such as the individual word, sentence, paragraph, section, or entire writing (e.g., a contract or statute).<sup>61</sup> However, this process can be complicated by the fact that many legal documents do not follow a strict template, as is the case for most judicial opinions. But even if the structure is relatively consistent across documents, choosing the right level of granularity requires balancing of competing factors. On one hand, if the segmentation is too coarse, algorithmic training is inefficient and more training data is required. In addition, modern language models simply do not work for very long text sequences, because their computational complexity increases with the length of the document at an exponential rate.<sup>62</sup> On the other hand, many more complex classification tasks require a sufficiently large context to give a complete and accurate answer. Therefore, if the segmentation is too granular, performance may suffer. Accurate segmenting can thus pose a significant challenge to NLP research.

When lawyers are working with legal texts, these challenges are exacerbated, because the most informative unit of analysis is often inconsistent and can vary

<sup>61</sup> KEVIN D. ASHLEY, *ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE* 277 (2017).

<sup>62</sup> See Ashish Vaswani et al., *Attention Is All You Need*, 2017 *ADVANCES IN NEURAL INFO. PROCESSING SYS.* 30. For an overview of potential solutions, see Chuhan Wu et al., *Fastformer: Additive Attention Can Be All You Need*, ARXIV (2021), <https://arxiv.org/abs/2108.09084>.

widely from one document to the next. To illustrate, assume that an algorithm is trained to automatically determine whether a court in a contracts dispute has personal jurisdiction over the defendant.<sup>63</sup> As input, we may provide a description of facts and the text of the agreement between the parties. However, choice-of-forum provisions can come in very different forms. Sometimes, they are contained in a single sentence. At other times, they may span multiple paragraphs. The agreement may even contain multiple dispute settlement provisions for different types of disputes. Thus, whether the appropriate unit of analysis is a sentence, paragraph, or section may vary from one document to the next, and this variation can significantly decrease classification performance and/or efficiency. To be sure, researchers have proposed deep-learning classifiers that are able to retain information as they move from one segment of the document to the next, thus ameliorating some of the concerns of inappropriate text segmentation.<sup>64</sup> However, because the computational complexity of these approaches increases exponentially in the length of the text sequence, they cannot feasibly be used to examine long legal documents.<sup>65</sup> Although researchers are beginning to examine ways around this limitation,<sup>66</sup> we are still far from finding workable solutions for long texts that obviate the need of segmentation.

In addition to document segmentation, identifying and processing the appropriate document structure complicates efforts to automate legal analysis. Many legal documents are highly contextual in that they rely on internal and external references in order to give meaning to their words or phrases. For example, M&A agreements typically include elaborate “Definitions” sections that define phrases such as “bankruptcy event” or “material adverse event,” which are then used throughout the contract. Similarly, regulations and statutes often rely on definitions that are not contained in the text of the document itself and thus can only be interpreted accurately by turning to the referenced text.<sup>67</sup> There may further be a hierarchical structure to legal documents, by which a definition only applies to parts of the document that are lower in the hierarchy. For instance, a definition contained in a statutory text may define words or phrases at the chapter level, but could then be modified by exceptions at the section level.

<sup>63</sup> Julian Nyarko, *We'll See You in ... Court! The Lack of Arbitration Clauses in International Commercial Contracts*, 58 INT'L REV. L. & ECON. 6, 6–24 (2019); Julian Nyarko, *Stickiness and Incomplete Contracts*, 88 U. CHICAGO L. REV. 1, 1–79 (2021).

<sup>64</sup> See, e.g., Zichao Yang et al., *Hierarchical Attention Networks for Document Classification*, 2016 PROC. N. AM. CHAPTER OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS 1480, 1480–89.

<sup>65</sup> The self-attention layer in a BERT model has a quadratic complexity of  $O(n^2)$  and thus, document length is generally limited to 512 tokens.

<sup>66</sup> Lulu Wan et al., *Long-Length Legal Document Classification*, ARXIV (2019), <https://arxiv.org/abs/1912.06905>.

<sup>67</sup> For instance, 2 C.F.R. § 175.25 (2021) defines a foreign public entity as “[a] public international organization, which is an organization entitled to enjoy privileges, exemptions, and immunities as an international organization under the International Organizations Immunities Act (22 U.S.C. 288–288f).”

Current language models are ill-equipped to recognize and appropriately process such structural idiosyncrasies with relevant precision. Although recent research is beginning to develop a promising methodology to try to accommodate legal document structure, the existing approaches are still highly domain-specific, and it is unclear whether they can be generalized.<sup>68</sup>

### 3.4.2 Availability of Training Data

A second key challenge is data availability. As discussed above, most ML models need access to large amounts of training data. In the context of legal tech applications based on NLP technology, this means access to large numbers of documents that are representative of the input data used in the application.

Whether data access is a potential problem depends on the nature of the legal tech application. One reason for this is that the creators and users of legal tech applications can create training data for some applications, while the same is not true for others. As a general matter, the creation of training data is usually possible when the “labels” that are to be predicted can be generated by the users of the legal tech application themselves. As an example, consider document review. For many document review tasks, training data can be created by tasking human coders (in many cases, lawyers and paralegals) with manually labeling a subset of documents according to a coding scheme that is later replicated by the ML algorithm.<sup>69</sup> By contrast, in outcome prediction tasks, labels (the outcomes of disputes) are the product of a complex interaction between a multitude of actors including judges and litigants. A party interested in predicting the outcome of a dispute cannot on its own generate additional cases to serve as training data. Rather, they are limited to the cases generated in the context of past disputes.

Creators of applications that depend on existing data often face the problem that suitable training data from past cases is not widely available. There are various reasons for this. One potential problem is that legal texts are oftentimes hidden behind paywalls (either erected by private database providers, for example Westlaw,<sup>70</sup> or public players such as PACER<sup>71</sup>). Even more important, however, is the problem that many relevant documents are never collected by any central entity that would be in a position to make them available to potential creators of legal tech applications. For example, while it would be desirable to create outcome prediction

<sup>68</sup> Dominic Seyler et al., *Finding Contextually Consistent Information Units in Legal Text* (2020) (unpublished manuscript).

<sup>69</sup> See also Remus, *The Uncertain Promise*, at 1701–1706.

<sup>70</sup> One example of a potential useful database maintained by a commercial database provider is Pleadings, Motions, and Memoranda on Westlaw Edge. See *Pleadings, Motions, and Memoranda*, THOMPSON REUTERS, <https://legal.thomsonreuters.com/en/products/westlaw/pleadings-motions-memoranda>.

<sup>71</sup> See Chapter 14 in this volume.

tools that can use early information about a dispute as input data, such information is almost never collected in a systematic way.

Selective availability of information is problematic not only because it raises questions about the availability of sufficiently sized training datasets to achieve high levels of predictive accuracy. It also gives rise to concerns about biased results. In the context of outcome prediction, because there are (virtually) no data repositories that systematically collect texts related to legal disputes prior to litigation, whether information about a dispute will become available at some point often depends on whether the dispute ends up in court or not. This is a problem because cases that do not result in litigation arguably differ in important ways from those that are litigated.<sup>72</sup> For instance, if there is clear, established precedent or otherwise little ambiguity in the rules that apply to a dispute, litigation is unlikely to occur. But this means that the available textual data is not an accurate representation of all potential disputes. Instead, textual data regarding disputes that end up in litigation will be heavily overrepresented. In the extreme, an outcome prediction tool that never encounters an “easy” case during the training process and is instead only trained on “hard” cases will have a tendency to abstain from making strong predictions even if a trained lawyer would be able to predict the outcome with certainty.

### 3.4.3 No Benchmark Data

The final limitation we discuss is the absence of domain-specific benchmark data that would allow creators of legal tech applications to compare the efficacy of newly developed methodologies. NLP is a research field that develops at a rapid pace. Indeed, the number of yearly submissions to ACL Anthology, the largest source of NLP papers, increased from about 1,000 in the 2000s to more than 4,000 new submissions by 2019.<sup>73</sup> In order to identify breakthrough contributions, the NLP community relies heavily on the concept of benchmarking. Benchmarks such as SuperGLUE<sup>74</sup> or the Stanford Question Answering Dataset (SQuAD)<sup>75</sup> contain hundreds of thousands of observations corresponding to various non-legal tasks, such as question answering, causal reasoning, and reading comprehension. The benchmarks have easily accessible leaderboards<sup>76</sup> and achieving state of the art (or SOTA) results is the main way in which important innovations are identified and adopted.

<sup>72</sup> George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J.L. STUD. 1 (1984).

<sup>73</sup> Saif M. Mohammad, *NLP Scholar: A Dataset for Examining the State of NLP Research*, 12 PROC. LANGUAGE RES. & EVALUATION CONF. 868, 868–77 (2020).

<sup>74</sup> Alex Wang et al., *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*, 33 PROC. INT'L CONF. ON NEURAL INFO. PROCESSING SYS. (2020).

<sup>75</sup> Pranav Rajpurkar et al., *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, ARXIV (2016), <https://arxiv.org/abs/1606.05250>.

<sup>76</sup> *Leaderboard Version: 2.0*, SUPERGLUE, <https://super.gluebenchmark.com/leaderboard/>; SQuAD2.0, SQuAD, <https://rajpurkar.github.io/SQuAD-explorer/>.

However, although these benchmarks aim to capture language understanding of NLP models, they are designed for an ordinary/natural use of language. As mentioned above, legal language differs from ordinary language in several important aspects. Whether a higher score at SuperGLUE or SQuAD corresponds to better performance at legal interpretation remains unknown – and, indeed, there are many reasons to doubt that the correlation is close to perfect. This suggests that we currently lack the means to gauge how well language models work in the legal domain. To remedy this problem would require the availability of benchmark datasets specifically designed to resemble legal language understanding and reasoning tasks. However, such datasets can be exceedingly resource-intensive to design and maintain.<sup>77</sup> Indeed, it appears likely that significant progress can be made only if there is a shared commitment toward advancing and improving the use of natural language processing in law. Naturally, there appear to be various barriers to such a collaborative process. For instance, it stands to reason that law firms would be in a particularly good position to make available representative samples of legal documents (such as contracts). At the same time, law firms generally lack incentives to make these documents available to a general audience for benchmarking. To be sure, there are some efforts to try to overcome these barriers. For instance, the Atticus Project has made publicly available a dataset of annotated contracts with the stated goal of establishing a reliable reference corpus that could be used to improve the use of AI in contract analysis.<sup>78</sup> However, with a corpus of 510 agreements, the dataset is still far from providing a representative sample of commercial contracting that is comparable to those available to benchmark natural language models. It remains to be seen whether efforts like the Atticus Project can establish themselves as reliable benchmarks in the legal domain.

### 3.5. CONCLUSION AND OUTLOOK

NLP, a set of computational techniques that automate the extraction and processing of information from unstructured text, is considered a key ingredient in many legal tech applications. However, the performance of these techniques might not always live up to some commentators' high expectations. In particular, as this chapter argues, current NLP techniques are ill-equipped to distill legal concepts from texts, which imposes severe limitations to their use in legal tech applications that need to perform the equivalent of legal reasoning. Further, creators of NLP-assisted legal

<sup>77</sup> One clever exception to this rule is presented in Lucia Zheng et al., *When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings*, 18 PROC. INT'L CONF. ON A.I. & L. 159, 159–68 (2021). By algorithmically extracting holdings of cases, they are able to automatically create multiple-choice questions that are then used to evaluate the performance of various language models.

<sup>78</sup> Dan Hendrycks et al., *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*, 35 PROC. INT'L CONF. ON NEURAL INFO. PROCESSING SYS. (2021).

tech need to grapple with several additional challenges, including a dearth of training data and the absence of benchmark datasets.

This result suggests that, in the short to medium term, NLP can be employed most fruitfully in tasks that do not significantly rely on legal reasoning. In contrast, the automation of tasks that rely heavily on legal reasoning will remain out of reach for the foreseeable future, save for any significant changes in how language models operate. This is because, for automated legal reasoning to succeed, it is not sufficient for an algorithm to process language in a literal sense. Instead, legal reasoning requires building, understanding, and processing the legal ontology that the language reflects. In other words, what is needed is a particularly complex form of Natural Language Understanding.<sup>79</sup>

How, then, might we identify tasks that are more reliant on legal reasoning and less amenable to automation? Perhaps the most obvious examples are those applications that require not just a high performance in predicting outcomes, but also need to explain how they arrived at the prediction.<sup>80</sup> But concerns about explainability are not the only reason why the ability of algorithms to perform the equivalent of legal reasoning matters. This ability is also a crucial determinant of an algorithm's predictive performance in those tasks for which large enough amounts of representative training data are not available.

If such training data is available, prediction tasks reduce to an inference problem: The algorithm must simply identify the predictors of a legal outcome while filtering out the noise. This can be the case, for instance, for outcome prediction tasks in run-of-the-mill cases that are based on relatively homogenous fact patterns.

However, if representative training data does not exist, accurate predictions are more heavily reliant on the availability of a representation of the general principles underlying previous case law. Without these, algorithms will likely fail to generate accurate predictions for previously unseen fact patterns. Because of current NLP techniques' limitations in recovering legal ontologies, NLP-based automation will likely face the greatest technological hurdles in dynamic areas of law where rules frequently change (e.g., those characterized by significant regulatory activity, such as environmental law), in areas that are multi-faceted and complex, and in areas in which the data is scarce (e.g., in litigation areas where settlement rates are particularly high). Similarly, settings that produce many new, previously unforeseen legal constellations will prove to be particularly challenging (e.g., appellate court decisions).

<sup>79</sup> Alex Wang et al., *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, ARXIV (2019), <https://arxiv.org/abs/1804.07461>.

<sup>80</sup> Engstrom & Gelbach, *Legal Tech*, at 1025.