Researching Legal AI: The Cambridge Law Corpus and Predicting Decisions of the UK Employment Tribunal

Holli Sargeant and Felix Steffek

ABSTRACT

This contribution introduces the Cambridge Law Corpus (CLC) and a research project benchmarking the prediction of UK Employment Tribunal decisions, which is based on the CLC data. The CLC is a dataset containing more than 320,000 UK court decisions. This article explains the need for legal datasets, the creation of the CLC and the ethical considerations concerning the dataset's construction and distribution. Subsequently, an experiment engaging with legal judgment prediction using the dataset is reported. The decisions predicted are those of the UK Employment Tribunal, which is the first instance for conflicts between employees and their employers. The experiment compares baselines of different AI models and human experts predicting whether the employee will win, partly win, lose or whether the Tribunal will render another decision.

1. THE CAMBRIDGE LAW CORPUS: A DATASET FOR LEGAL AI RESEARCH

The Cambridge Law Corpus (**CLC**) represents a groundbreaking advancement for legal AI research in the UK. We present the first and only large-scale dataset of machine readable UK court cases for computational research. This dataset of over 320,000 UK court cases spans from the 16th century to the present, with most cases originating in the late 20th and early 21st centuries. The CLC establishes the research infrastructure required to advance legal AI research traditionally hindered by access to large-scale, structured legal data. It has been created by an interdisciplinary team, consisting of Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson and Felix Steffek. The paper introducing the CLC has been published by Advances in Neural Information Processing Systems 36 (NeurIPS 2023): Datasets and Benchmarks Track.[1]

Recent advancements in AI and natural language processing (NLP) have been remarkable, especially with the development of transformer-based models like BERT and large language models such as GPT. These models have achieved or even surpassed human performance in various language tasks. While their application to the legal domain is a rapidly developing area, it is limited by the scarcity of specialised legal datasets. One of the primary strategies for enhancing the capabilities of legal AI involves pre-training language models. Therefore, legal AI development hinges substantially on the availability and quality of legal data, which is distinct from general corpora. First, case law contains complex, nuanced, and domain-specific language. Second, it is jurisdiction-specific, making it challenging to develop models that are specific to different legal systems. Third, the inherent lack of metadata or structure in UK case law further complicates the application of AI, which thrives on large, well-structured data.

The CLC aims to bridge this gap by providing a rich, structured dataset tailored for legal AI research. It currently contains case law from 53 courts and tribunals across the UK, particularly focusing on England and Wales. It is continuously updated, for example, judgments from Scotland and Northern Ireland will be added in due course. The dataset is organised by court and year, where each case is stored as a single XML file containing the legal text and certain metadata including an assigned unique identifier (CLC-ID) and neutral citation. Additionally, we include a small set of expert annotations for case outcomes to assist advanced research tasks like outcome prediction and extraction. Using our annotated data, we have trained and evaluated case outcome extraction with GPT-3.5, GPT-4 and RoBERTa models to provide benchmarks for future research.

1 Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson and Felix Steffek, *The Cambridge Law Corpus: A Dataset for Legal AI Research,* Advances in Neural Information Processing Systems 36 (NeurIPS 2023): Datasets and Benchmarks Track, available at <https://papers.nips.cc/paper_files/paper/2023/hash/819b8452be7d6af1351d4c4f9cbdbd9b-Abstract-Datasets_and_Benchmarks.html>.

The CLC can be used for diverse research tasks and applications; we consider two in our paper. Case outcome extraction, for example, allows models to locate judgment outcomes within lengthy documents, a challenging task well-suited to automation. In early experiments, transformer-based models and large language models show differing levels of accuracy in identifying outcome-related information. Another example for computational analysis about case law includes topic modelling. This research enables analysis of long-term trends in legal areas, such as contract disputes and employment law, shedding light on the evolving factors influencing UK court decisions and access to the legal system. The CLC also opens up a multitude of research opportunities in the field of legal AI and broader computational analysis of law. By providing a comprehensive and structured dataset, the CLC provides the research infrastructure to explore such opportunities.

The legality and ethics of collecting, processing and releasing the corpus is of paramount importance. We have undertaken considerable analysis of the relevant considerations for lawful and ethical design of this project. One core concern with the release of large legal datasets is the personal information they contain. To uphold principles of open justice, UK court cases are generally not anonymised. However, where necessary for the proper administration of justice or to protect certain parties—such as children, victims of sexual offences or asylum seekers—the court will anonymise identities. Privacy regulations, specifically the Data Protection Act 2018 and UK implementation of the European Union's General Data Protection Regulation, detail how personal data can be handled. We have prioritised the use of this corpus in a way that is in the public interest and does not pose risks to individuals' rights, freedoms or interests. By balancing the public availability of all cases in the dataset in other repositories and the principle of open justice, with our prohibition of research identifying individuals, the requirement of ethical clearance and our mechanisms for the erasure of data, we believe these are appropriate safeguards to avoid harm to any individuals.

Against this background, the CLC is not open access. Only researchers can gain access through a straightforward application form.[2] We ask that university-affiliated researchers provide a research plan, university ethical approval and agree to the Terms and Conditions. These requirements help ensure the corpus is used responsibly, aligning with UK laws and ethical research standards.

The CLC has established critical infrastructure for legal AI research in the UK. We are committed to the continuous improvement of the CLC. Future updates will include additional cases, enhanced annotations, and new features based on user feedback and emerging research needs. As more researchers engage with this corpus, the opportunities for impactful insights and transformative advancements in legal AI will continue to expand, reshaping the future of legal research and accessibility.

The work on the CLC is part of the UK Economic and Social Research Council (ESRC) and JST (Japan Science and Technology Agency) funded project on Legal Systems and Artificial Intelligence. The support of the ESRC and JST is gratefully acknowledged.

2. BENCHMARKING CASE OUTCOME PREDICTION FOR THE UK EMPLOYMENT TRIBUNAL: THE CLC-UKET DATASET

Employment tribunals play a critical role in resolving disputes between employers and employees, yet the volume and complexity of cases create challenges for timely and consistent resolution. Predicting case outcomes through advanced AI can enhance access to justice, streamline legal processes and help stakeholders make better-informed decisions. In a recent paper published by the Association for Computational Linguistics in the Proceedings of the Natural Legal Language Processing Workshop 2024, Huiyuan Xie, Felix Steffek, Joana Ribeiro de Faria, Christine Carter and Jonathan Rutherford explore the intersection of technological innovation and access to justice, focusing on the development of benchmarks for predicting case outcomes within the UK Employment Tribunal (**UKET**).[3]

Despite the potential benefits of predictive models in legal contexts, there remains a notable gap in available legal data that hampers AI advancements. Publicly accessible, comprehensive datasets are rare, particularly those that offer standardised annotations of legal decisions. Addressing this gap, the CLC-UKET dataset created as part of this project offers a solution by providing an extensive, curated collection of UKET cases, annotated and organised to enhance predictability and transparency within employment dispute resolution.

The CLC-UKET dataset was curated from the Cambridge Law Corpus,[4] compiling approximately 19,000 UKET cases. The dataset includes intricate legal annotations across multiple facets, making it a comprehensive resource for legal AI applications. Manual annotation by legal experts is a time-consuming and costly process. To alleviate this burden, we explored the use of large language models (**LLMs**) to automate the annotation process. By utilising LLMs, specifically the GPT-4-turbo model, we efficiently handled vast quantities of data without compromising on the accuracy or depth of information. Through an iterative approach to prompt design, we

2 The application form and associated information are available at <https://www.cst.cam.ac.uk/research/srg/projects/law>.

3 Huiyuan Xie, Felix Steffek, Joana De Faria, Christine Carter and Jonathan Rutherford, *The CLC-UKET Dataset: Benchmarking Case Outcome Prediction for the UK Employment Tribunal*, Proceedings of the Natural Legal Language Processing Workshop 2024, pp. 81–96, available at <https://aclanthology.org/2024.nllp-1.7/>.

4 Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson and Felix Steffek, *The Cambridge Law Corpus: A Dataset for Legal AI Research*, Advances in Neural Information Processing Systems 36 (NeurIPS 2023): Datasets and Benchmarks Track, available at <https://papers.nips.cc/paper_files/paper/2023/hash/819b8452be7d6af1351d4c4f9cbdbd9b-Abstract-Datasets_and_Benchmarks.html>.

optimized the LLM's performance for annotating the following details: (1) facts, (2) claims, (3) references to legal statutes, (4) references to precedents, (5) general case outcomes, (6) general case outcomes labelled as "claimant wins", "claimant loses", "claimant partly wins", and "other", (7) detailed orders and remedies and (8) reasons. We report on this process in more detail in another paper available on SSRN and arXiv.[5]

The annotated CLC-UKET dataset allows for case outcome prediction, a challenging but valuable task in legal AI. Acknowledging discussion on task terminology,[6] we use the term "prediction" rather than "classification" because we specifically focus on predicting case outcomes using only facts and claims, without including explicit outcome information in the input data. In this prediction task, given a set of case facts and claims, the model generates an outcome label that falls into one of four categories: "claimant wins", "claimant loses", "claimant partly wins" or "other". This task relies solely on the description of facts and claims, intentionally excluding any explicit details about the tribunal's final decision to test the model's predictive capabilities based on input case summaries alone. To establish a baseline for model performance, human predictions were collected by providing experts access to the same facts and claims without the actual case outcomes. Comparing human predictions to model outputs is crucial for understanding the limitations and strengths of AI in this domain.

Four types of approaches were used to benchmark the dataset's predictive potential. Each type offers a unique approach, and their comparative performances shed light on the effectiveness of model customisation for complex legal tasks.

1. Performance of Finetuned Transformer Models

- **Highest F-Scores Overall**: Among all models, **finetuned transformer models**, particularly T5, achieved the best results, showing superior accuracy in predicting outcomes. The T5 model displayed the highest F-scores across most categories, highlighting the advantage of training models specifically on the CLC-UKET dataset.
- **Precision and Recall Strengths**: The T5 model achieved strong precision and recall scores across the categories of "claimant wins" and "claimant loses." For instance, T5 attained an F-score of **0.650 for "claimant wins"** and **0.734 for "claimant loses"**. This accuracy underscores how model fine-tuning on specific legal annotations can enhance precision in interpreting complex tribunal judgments.
- **Gaps in Specific Categories**: Despite its overall performance, the T5 model struggled with the categories "claimant partly wins" and "other", where it achieved low F-scores. The "other" category in particular yielded an F-score of zero, suggesting that even advanced models face challenges with underrepresented or very complex outcomes. This outcome indicates that finer distinctions in nuanced cases may require additional tailored training or refined annotation strategies.

2. Comparative Analysis of GPT-3.5 and GPT-4 Models

- **Small but Notable Improvements with GPT-4**: Between the two GPT-based models, **GPT-4 consistently outperformed GPT-3.5**, although the margin was relatively small. This improvement highlights the incremental advancements in newer LLM versions and how refined language models contribute to higher accuracy in complex legal tasks.
- **Impact of Few-Shot Examples on GPT-3.5's Accuracy**: Interestingly, incorporating **task-specific few-shot examples** significantly enhanced GPT-3.5's performance. For instance, using few-shot examples that matched the legal area of the target case improved its F-score in outcome prediction more effectively than randomly sampled examples. This result emphasises the importance of contextual relevance when leveraging few-shot learning, especially in specialised fields like legal AI where case-specific nuances matter.
- **GPT-4 Zero-Shot Precision**: Notably, **GPT-4 achieved the highest precision in its zero-shot setting** among all baseline models, indicating that it can accurately predict outcomes without task-specific fine-tuning when given the right context. Providing task-related examples in few-shot settings (specifically the **"juris-2"** setting, where two examples from similar legal areas were provided) boosted GPT-4's F-score. However, the relatively modest gains suggest that simply adding more examples does not drastically improve performance, pointing to a need for high-quality, highly relevant few-shot examples.

3. Benchmarking Against Human Expert Predictions

- **Human Predictions Outperform AI**: A critical reference point for the model's efficacy was **human expert predictions**, which outperformed the AI models by an approximately **19% higher F-score** over the best-performing model, T5. This gap highlights

5 Joana Ribeiro de Faria, Huiyuan Xie and Felix Steffek, *Automatic Information Extraction for Employment Tribunal Judgements Using Large Language Models*, available at <https://ssrn.com/abstract=4776160> and <https://arxiv.org/abs/2403.12936>, submitted to journal.

6 Masha Medvedeva and Pauline McBride, *Legal Judgment Prediction: If You Are Going to Do It, Do It Right*, Proceedings of the Natural Legal Language Processing Workshop 2023, pp. 73–84, available at <https://aclanthology.org/2023.nllp-1.9/>.

the value of human expertise in interpreting legal nuances that current AI models struggle to replicate.

- **Strength in Judgment-based Decisions**: Human expert annotators demonstrated the highest F-scores for both "claimant wins" and "claimant loses" categories, indicating that the subjective analysis of case nuances may require human interpretation that AI has yet to achieve. On the other hand, GPT-4 outperformed the human experts when predicting "claimant partly wins" and "other", i.e., in more complex cases.

4. Benchmarking Hard Cases

- **Predicting Hard Cases: The human experts were asked to identify those cases that they considered as hard to predict. This allowed comparing the models' and human performance as regards hard cases. As expected, both AI models and human experts achieved worse scores for hard cases.**
- **Finetuned Transformer Models are Best in Predicting Hard Cases: Interestingly, the finetuned transformer models, in particular T5, outperformed both the GPT-based models and the human experts in predicting hard cases.**

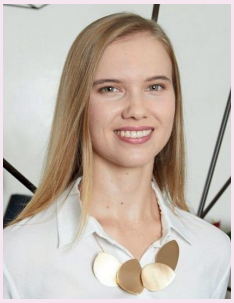
Whilst the study provides valuable insights into the prediction of dispute outcomes for the UK Employment Tribunal, it is important to acknowledge certain limitations of our findings. First, information leakage, one example of bias in legal data,[7] may arise from using LLM summaries of judge written case judgments as we are unable to use neutral information. This information might reflect the judges' post-hoc knowledge and subjective perspectives that shape their written judgments and any information leakage from the LLM summary. Second, while GPT-4 was used for efficient annotation, automated extraction may contain minor inaccuracies, and more detailed factual data could improve predictions. Finally, the dataset spans 2011-2023, during which legal rules and principles evolved, possibly affecting model accuracy over time, as decision dates were indirectly inferred. Future research will address these aspects for more robust prediction models.

The CLC-UKET dataset establishes a meaningful benchmark in legal AI, offering a robust resource for advancing outcome prediction in employment tribunals. Access to the CLC-UKET dataset is available through the Cambridge Law Corpus.[8] While AI models demonstrate promising accuracy, particularly with fine-tuning, human expertise still outshines AI in relevant areas. As we move forward, exploring ways to bridge this gap and improve AI's adaptability will be key to realizing a future where predictive AI and human judgment work seamlessly to enhance access to justice.

This project received funding support from the Cambridge Centre for Data-Driven Discovery and Accelerate Programme for Scientific Discovery, made possible by a donation from Schmidt Futures.

7 Holli Sargeant and Måns Magnusson, *Bias in Legal Data for Generative AI*, 2nd Workshop on Generative AI and Law (GenLaw '24), available at <https://icml.cc/virtual/2024/39169>.

8 At <https://www.cst.cam.ac.uk/research/srg/projects/law>.



Holli Sargeant

Holli Sargeant is a PhD Candidate in the Faculty of Law, University of Cambridge funded by the General Sir John Monash Foundation. Her research examines the consequences of algorithmic decision-making. Her research focuses on the intersection of artificial intelligence and law, exploring two key areas: the necessary adaptations of legal frameworks to address AI-related risks, and the potential applications of AI to the legal system itself. Holli works with various international organisations and not-for-profits to provide legal advice on the use of emerging technology to improve access to justice and uphold human rights.

Prior to commencing her PhD, Holli was an Australian solicitor working in digital law, technology transactions and human rights at Herbert Smith Freehills and the Australian Human Rights Commission. Holli holds a Bachelor of Laws with first class Honours and a Bachelor of International Relations from Bond University. She has previously studied at the National University of Singapore and worked at the Singapore Academy of Law as a New Colombo Plan Scholar.

https://www.law.cam.ac.uk/people/research-students/h-sargeant/79151.

Felix Steffek

Felix Steffek is Professor of Law at the University of Cambridge and Senior Member of Newnham College. He serves as Co-Director of the Centre for Corporate and Commercial Law (3CL) and holds a JM Keynes Fellowship in Financial Economics awarded by the University of Cambridge. He is Global Distinguished Professor of Law at the University of Notre Dame.

His research interests cover corporate finance and insolvency law, artificial intelligence, dispute resolution and commercial law. He has advised international organisations, governments, parliaments and courts in these areas. He represents law on the Academic Publishing Committee of Cambridge University Press.

Felix Steffek is leading multiple research projects on artificial intelligence and law, among them the Nuffield Foundation funded project on 'Access to Justice Through Artificial Intelligence' and the AHRC funded project on 'Explainable and Ethical Legal Artificial Intelligence'.

For further information please see https://www.law.cam.ac.uk/people/academic/f-steffek/6136.