How well do SOTA legal reasoning models support abductive reasoning?*

Ha Thanh Nguyen[1,*], Randy Goebel[2], Francesca Toni[3], Kostas Stathis[4] and Ken Satoh[1]

[1]National Institute of Informatics (NII), 2-1-2 Hitotsubashi, Chiyoda City, Tokyo, Japan
[2]University of Alberta, 116 St & 85 Ave, Edmonton, AB T6G 2R3, Canada
[3]Imperial College London, Exhibition Rd, South Kensington, London SW7 2BX, United Kingdom
[4]Royal Holloway University of London, Egham Hill, Egham TW20 0EX, United Kingdom

Abstract

We examine how well the state-of-the-art (SOTA) models used in legal reasoning support abductive reasoning tasks. Abductive reasoning is a form of logical inference in which a hypothesis is formulated from a set of observations, and that hypothesis is used to explain the observations. The ability to formulate such hypotheses is important for lawyers and legal scholars as it helps them articulate logical arguments, interpret laws, and develop legal theories. Our motivation is to consider the belief that deep learning models, especially large language models (LLMs), will soon replace lawyers because they perform well on tasks related to legal text processing. But to do so, we believe, requires some form of abductive hypothesis formation. In other words, while LLMs become more popular and powerful, we want to investigate their capacity for abductive reasoning. To pursue this goal, we start by building a logic-augmented dataset for abductive reasoning with 498,697 samples and then use it to evaluate the performance of a SOTA model in the legal field. Our experimental results show that although these models can perform well on tasks related to some aspects of legal text processing, they still fall short in supporting abductive reasoning tasks.

Keywords

neural networks, abductive reasoning, legal reasoning

1. Introduction

The rise of transformer-based deep learning models [1] has brought remarkable advancements in natural language processing (NLP), including tasks related to legal text processing [2, 3, 4, 5, 6, 7]. These advancements have the potential to improve access to justice and legal services for underserved communities, and to enhance the efficiency and accuracy of judicial processes.

However, a critical component of legal intelligence is abductive reasoning, which is of paramount importance for lawyers and legal scholars in formulating logical arguments, interpreting laws, and developing legal theories [9, 10]. Figure 1 is an example of statute law retrieval task requiring abductive reasoning skill in the existing COLIEE Competition [8]. Since

London'23: Workshop on Logic Programming and Legal Reasoning, July 09–10, 2023, London, UK
*Corresponding author.
nguyenhathanh@nii.ac.jp (H. T. Nguyen); rgoebel@ualberta.ca (R. Goebel); f.toni@imperial.ac.uk (F. Toni); Kostas.Stathis@rhul.ac.uk (K. Stathis); ksatoh@nii.ac.jp (K. Satoh)
0000-0003-2794-7010 (H. T. Nguyen); 0000-0002-0739-2946 (R. Goebel); 0000-0001-8194-1459 (F. Toni); 0000-0002-9946-4037 (K. Stathis); 0000-0002-9309-4602 (K. Satoh)

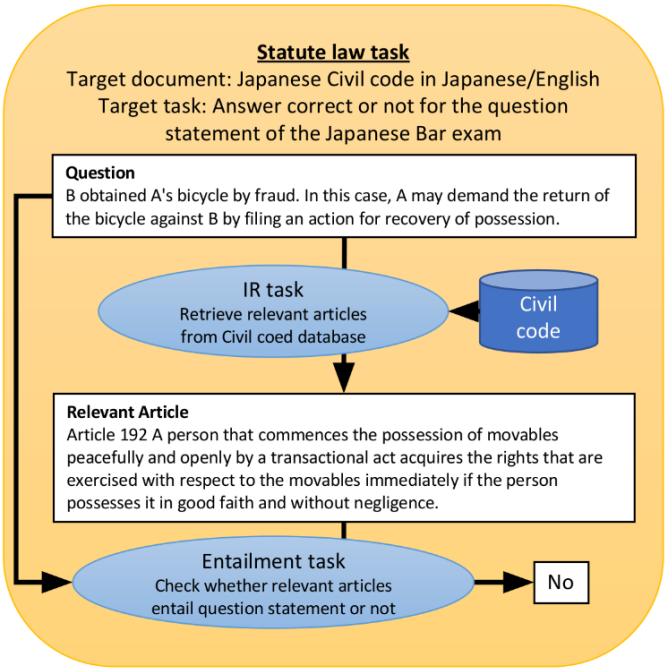

Figure 1: An example of statute law retrieval task requiring abductive reasoning skill in COLIEE Competition [8]. The truth value (No) of the statement is determined by assuming that the statement is either true or false, and then searching for the appropriate legal article that provides the best explanation for that assumption. In this case, the relevant article is Article 192, which provides an exception to A's ability to demand the return of the bicycle against B through an action for recovery of possession. Abductive reasoning is utilized in determining that the statement's value is "No" based on the application of Article 192 and the exception it presents.

transformer-based models have shown promising results in various NLP tasks within the legal field, it is essential to evaluate their performance on tasks involving abductive reasoning to better understand their capabilities and limitations as AI tools in the legal field.

Existing literature has explored the development and evaluation of transformer-based models for legal text processing tasks such as legal document retrieval, summarization, entailment, and question-answering. Yet, there is a gap in evaluating these models for the crucial task of abductive reasoning. To address this gap, we first articulate an empirical method to reveal the role of abductive reasoning in legal information processing. We identify an appropriate corpus of examples by examining the ART dataset [11], which is the first large dataset for abductive reasoning tasks, and αNLI task, which investigates the viability of language-based abductive reasoning.

In this paper, we focus on:

- Enhancing the reliability of the dataset for abductive reasoning through task redefinition and data augmentation.
- Evaluating the performance of a state-of-the-art (SOTA) transformer-based model in the legal field on abductive reasoning tasks.

Our experimental results show that although the selected SOTA model can perform well on tasks related to legal text processing, it still falls short in supporting abductive reasoning tasks, shedding light on an important limitation of these models in legal reasoning. This study provides a more comprehensive understanding of the capabilities of transformer-based models in the legal domain, emphasizing the importance of abductive reasoning, which is often overlooked in related research.

2. Background

2.1. Legal Text Processing

There has been significant research on the use of artificial intelligence and machine learning (AI) in the legal field in recent years. This research has resulted in the development of a number of state-of-the-art models that are able to perform various tasks related to legal text processing, such as contract risk analysis and case law retrieval.

Deep learning models have the ability of automated latent feature extraction, which allows us to use these models not only for similarity matching tasks but also in other semantic matching tasks such as question answering [12, 13], machine reading comprehension [14, 15], image retrieval [16] and entity matching [17, 18]. These tasks are all important for the general challenge of legal reasoning.

In the legal retrieval task, legal documents are usually structured and existing systems are designed to retrieve relevant legal texts (e.g., regulations) based on a given query. This task is an essential component of intelligent legal counsel systems and commonly appears in legal automated processing competitions [19, 8, 20, 21]. One of the challenges in legal information retrieval is that the available data is usually very limited. This is why the current best systems often need to be based on some supportive rules or data augmentation methods. As a consequence, Deep learning with transfer learning methods has been successfully applied to this problem in a number of ways [12, 2, 22, 23, 24].

Retrieval tasks are foundational to many other legal text processing tasks. For example, contract analysis often involves retrieving relevant provisions or clauses from a contract based on a given query. Similarly, case law retrieval involves identifying and retrieving relevant case law and legal precedent based on a given query. These tasks require making inferences and educated guesses about what information is likely to be relevant, based on the characteristics of the query and the available data.

In addition, many other legal text-processing tasks, such as legal document classification and summarization, are also related to retrieval. For instance, classifying a legal document as relevant or irrelevant to a given case may involve retrieving similar documents and using them as a reference. Similarly, summarizing a legal document may involve retrieving relevant information and condensing it to a shorter form. Overall, retrieval tasks form the foundation for many legal text processing tasks, and the ability to perform retrieval effectively is essential for the development of intelligent legal counsel systems and other AI tools for the legal field.

Note that retrieval tasks and abductive reasoning are closely related. In retrieval tasks, the goal is to identify and retrieve relevant information from a database or collection of documents based on a given query. This requires making an educated guess or inference about what

information is likely to be relevant, based on the characteristics of the query and the available data. Similarly, in abductive reasoning, the goal is to construct an argument or form a hypothesis based on a set of observations and a limited amount of information. This also requires making an educated guess or inference based on the available evidence, in order to explain the observations and arrive at a plausible conclusion.

2.2. Abductive Reasoning

Our approach to test deep learning models based on transformers [1] for reasoning is focused on abductive reasoning tasks, because abductive reasoning is such an important aspect of human reasoning [25]. In a typical abductive reasoning problem, one is given a set of observations, and the goal is to identify a hypothesis that can best explain the observations. This process can be divided into two steps: 1) potential hypothesis identification, and 2) hypothesis evaluation. In the first step, we need to identify a small set of potential hypotheses that are likely to explain the observations. In the second step, we need to evaluate the potential hypotheses and rank them within the current context of use, e.g., assess a set of alternative hypotheses w.r.t. which provides the basis for a "best" explanation.

Because the complexity of the real world is high, it is impossible to consider all potential hypotheses. Instead, we typically assume that a set of potential hypotheses is given, and the evaluation of the hypotheses is based on logical reasoning. This second step is similar to the scoring or ranking function used in the second step of an information retrieval process.

There are several ways to limit the number of potential hypotheses, for example, limiting the context, constraining the search space, or using a heuristic search method. Generally, only commonsense heuristic reasoning can help to limit the number of potential explanations. For example, suppose we want to explain why a health professional regulator received a complaint. In that case, we are only interested in explanations that are related to the health and care professionals' conduct or practice according to the relevant regulations [26, 27]. Therefore, to be useful in application, we need a mechanism to reduce the set of possible explanations.

Bhagavatula et al. [11] introduce the ART dataset, which contains 20K commonsense narrative contexts and 200K explanations. They approach abductive reasoning as the problem of finding the hidden middle of a linear series of the form, $\mathcal{O}_1 \wedge \mathcal{H} \implies \mathcal{O}_2$, where we observe $\mathcal{O}_2$, and $\mathcal{O}_1$, and we try to come up with the best hypothesis $\mathcal{H}$. In one example, they show two observations:

- **Observation 1:** Jenny cleaned her house and went to work, leaving the window just a crack open.
- **Observation 2:** When Jenny returned home she saw that her house was a mess!

In *αNLI*, the task of a model is to choose the most plausible explanatory hypothesis among the given candidates. For example:

- **Hypothesis 1:** A thief broke into the house by pulling open the window.
- **Hypothesis 2:** At work, she opened her window and the wind blew her papers everywhere.

Of the hypotheses above, if their common sense semantics are accepted, Hypothesis 1 is the most plausible for the two given observations. However, if we consider the possible existence of enthymemes, the outcome may be different. For example, if "the house" in Hypothesis 1 does not refer to Jenny's house, then Hypothesis 1 becomes completely inappropriate as an explanation for the observations. The existence of enthymemes affects the reliability of ART and αNLI for two reasons:

1. Dataset ART is constructed by crowd-sourcing with the fact that different people will have different sets of implicit arguments to support their different decisions.
2. As a consequence, in αNLI task, the model simply chooses a hypothesis which is more plausible than the other. That is, the model always outputs a **Yes** label and a **No** label for a pair of inputs, it does not happen that both hypotheses are plausible (or both implausible). In addition, within this setting, annotators may introduce their bias into the training and evaluation of the model.

3. Dataset Construction

While crowdsourcing can be an effective way to build a large dataset for evaluating the performance of deep learning models on tasks which require some logical capability [11], such as abductive reasoning, it is important to carefully verify the soundness and quality of the data. This can help to ensure that the dataset accurately represents the task and does not contain errors or biases that could impact the model's performance.

One way to increase the number of data points and ensure the soundness of the data is to use logic-based data generation techniques, such as symbolic reasoning or logical theorem proving. This can help to generate a larger number of high-quality examples of abductive reasoning that are consistent with the intended task and application domain.

Additionally, it is important to define the task clearly and accurately based on the characteristics of the data. If the task is not well-defined or does not align with the characteristics of the data, the model's performance may be difficult to interpret or may be difficult to explain. Carefully defining the task and choosing a suitable evaluation metric can help to ensure that the model's performance is accurately assessed and interpreted.

First, we precisely formulate the problem. With two observations $\mathcal{O}_1$, $\mathcal{O}_2$, and a hypothesis $\mathcal{H}$, the model needs to verify the validity of the Expression 1.

$$\mathcal{O}_1 \wedge \mathcal{H} \implies \mathcal{O}_2 \tag{1}$$

We then analyze the two drawbacks of ART and subsequently construct L'ART as an expanded and improved version based on negation rules and theory generators.

3.1. Logical Consistency

Compared with conventional programming languages, natural language have higher semantic tolerance, but therefore lower logical consistency. This is why, until recently, direct natural language-to-software conversion tools have relatively limited use. This well-known challenge of natural language use can also be an issue affecting the quality of crowdsourced datasets

like ART and the models trained on it. Sharing the same view on this issue, [28] train their Transformer model with data generated from a logic-based program called a theorem generator. Expanding on this result, Gaskell et al. [29] introduce an adversarial framework to improve the logical consistency of these "soft" theorem provers. This is done by training a discriminator, which is then used to detect incorrect outputs from the theorem prover. The crucial insight of this work is that a model trained with improved logical consistency can be applied to the task of soft theorem-proving with higher accuracy.

Without the support of a logic-based program in the data generation process, ART does not provide any guarantee of logical consistency of its content. For example, there are some samples in the dataset whose plausibility determination is based heavily on the enthymeme in the evaluator's knowledge base, which has the potential to introduce inconsistency. For example, here is a sample from ART:

- **Observation 1:** Ron started his new job as a landscaper today.
- **Observation 2:** Ron is immediately fired for insubordination.
- **Hypothesis 1:** Ron ignores his boss's orders and called him an idiot.
- **Hypothesis 2:** Ron's boss called him an idiot.

In other words, without a logic-based program in place, the data generated by ART may not be completely consistent or accurate. Additionally, the evaluator's own biases and knowledge can influence the plausibility of certain samples, leading to inconsistencies in the dataset. The example above, where Ron is fired for insubordination, produces two different hypotheses for why that might be the case, which further illustrates this potential for inconsistency in the data generated by ART.

To overcome this drawback, instead of only requiring the model to choose between two given candidate hypotheses, we reformatted the dataset and forced the powerful pretrained models to *predict* the binary label without limiting the number of candidates. The negative samples are derived from the positive ones by using logical negation, which guarantees the positive triples contain the best hypotheses explaining the given observations. This is an important distinction from the way the dataset was used in the original setting. This adjustment can provide the ability to achieve logical consistency, and determine whether or not a candidate can be a valid hypothesis with the two given observations.

3.2. Observation-Hypothesis Interchangeability

From $\mathcal{O}_1 \wedge \mathcal{H} \implies \mathcal{O}_2$, we can deduce $\mathcal{H} \wedge \mathcal{O}_1 \implies \mathcal{O}_2$. In other words, the first observation and the hypothesis are interchangeable. More specifically, they are two events producing the second observation. Of the two events becomes the observation, while the event *we do not observe* becomes the hypothesis. However, in terms of logic, the two events hold equal footing, as they are both postulated to explain the second observation. From this, in terms of dataset construction, we can double the number of positive samples by reversing the role of the hypothesis and the first observation.

This logical reformulation helps us to realize that if we can not interchange the hypothesis and the first observation in a triple, the triple is not a valid abductive reasoning sample. For example, assume that we have a silly triple of $\mathcal{O}_1$, $\mathcal{H}$ and $\mathcal{O}_2$ as follows:

- $\mathcal{O}_1$: John is the smartest person in the class.
- $\mathcal{H}$: Every smart person has a green car.
- $\mathcal{O}_2$: John has a green car.

In this case, we cannot interchange the $\mathcal{H}$ and $\mathcal{O}_1$ to get:

- $\mathcal{O}_1$: Every smart person has a green car.
- $\mathcal{H}$: John is the smartest person in the class.
- $\mathcal{O}_2$: John has a green car.

This is because the inference chain requires one more piece of information (i.e., the argument for $\mathcal{O}_2$ from $\mathcal{O}_1$ and $\mathcal{H}$ is an enthymeme), namely: "The smartest person in the class is a smart person," which is not included in the triple. In the latter triple, the hypothesis "John is the smartest person in the class." is not reasonable given only the two observations "Every smart person has a green car." and "John has a green car."

3.3. Logic-augmented Dataset

In our dataset construction process, we consider the above two reformulation factors which are not considered in ART: (1) We use a logic-based theorem generator to ensure logical consistency in the data; (2) We use logical formulas to ensure the validity of the triples in terms of abductive reasoning. Based on these two transformations, we expand and improve the ART dataset and propose a new dataset called the *logic-augmented abductive reasoning dataset* (L'ART). L'ART is introduced as a dataset for a binary classification problem, where the model is trained to predict the validity of each provided triple.

As described in Section 2.2, in the ART dataset there are triples with high plausibility and others with low plausibility. We select the highly plausible triples as positive samples. With the logic-based theorem generator, we randomly generate positive samples that are logically consistent and find the inference chains which have at least two inference steps to extract $\mathcal{O}_1$, $\mathcal{H}$ and $\mathcal{O}_2$. The hypotheses are then reversed to double the number of positive samples.

Producing negative samples in the context of abductive reasoning is more challenging than that in ordinary negation. We are looking for a hypothesis to explain the given observation; but we can not apply a random strategy to generate negative samples as there is no way for us to know whether a random hypothesis is reasonable for the given observations. We therefore limit the possibilities through our strategy of exploiting negation, but applying this strategy is not straightforward in the context of abductive reasoning. Consider that randomly negating the operators in the Expression 1 might still yield a triple labeled as true. We handle this issue by first constructing a truth table for the Expression 1, and then uniformly negating the operators. The truth table for the Expression 1 is as in Table 1.

The first row of Table 1 corresponds to the truth values in the case of positive samples. We can easily see that the only logical option to negate Expression 1 is to negate the operator $\mathcal{O}_2$. Interestingly, when we interchange $\mathcal{O}_1$ and $\mathcal{H}$, the soundness of the system is not affected. This is because the soundness of the system is based on the logical equivalence of $\mathcal{O}_1 \wedge \mathcal{H} \implies \mathcal{O}_2$ and $\mathcal{H} \wedge \mathcal{O}_1 \implies \mathcal{O}_2$.

Table 1
Truth table for Expression 1: $\mathscr{O}_1 \wedge \mathscr{H} \implies \mathscr{O}_2$. The first row corresponds to the truth values in the case of positive samples.

$\mathscr{O}_1$	$\mathscr{H}$	$\mathscr{O}_2$	$\mathscr{O}_1 \wedge \mathscr{H} \implies \mathscr{O}_2$
T	**T**	**T**	**T**
F	F	F	T
F	F	T	T
T	F	F	T
T	F	T	T
F	T	F	T
F	T	T	T
T	T	F	F

Compared to ART, the data format, number of samples, and their logic consistency in L'ART are significantly improved. ART, for the positive (plausible) hypotheses, presents a narrative context to Amazon Mechanical Turk workers who were then asked to make assumptions and write natural language hypotheses for the two given observations. For the negative (implausible) hypotheses, the plausible hypothesis can be modified through minimal edits (up to 5 words). For the positive (plausible) hypotheses, we use a logic-based theorem generator to randomly generate positive samples that are logically consistent, and identify the inference chains which have at least two inference steps to extract $\mathscr{O}_1$, $\mathscr{H}$ and $\mathscr{O}_2$. We also reuse the positive samples from ART and reverse the role of the hypothesis and the first observation. For the negative (implausible) hypotheses, we use a truth table to construct negative samples. The truth table shows that to negate the expression $\mathscr{O}_1 \wedge \mathscr{H} \implies \mathscr{O}_2$, we need to change the value of $\mathscr{O}_2$ to false. The model is trained to predict the validity of each provided triple. The L'ART dataset has almost 2.5 times as many samples as the ART dataset, which contains 200k samples. So 476,167 of the samples in L'ART are used for training, 9,339 for validation, and 13,191 for testing. This dataset can be used as a benchmark for measuring the abductive reasoning skills of state-of-the-art models in the legal domain, but it's not limited to this purpose, as it can also be a valuable resource for other natural language processing and machine learning tasks, especially to consider complex NLP tasks because of transformers lack of reasoning ability.

4. Task Redefinition

In addition to the L'ART data's quantity and quality, task definition is also important in training and evaluating the model. Bhagavatula et al. [11] introduce the αNLI task, in a way that the model needs to select the most plausible explanatory hypothesis between the two given. We argue that although this task is appropriate for evaluating the ability of the model to perform abductive reasoning and find the most plausible explanation, the way the authors limit the αNLI task to a binary classification of two hypotheses makes the problem easier and can lead to an overfitting of the model. The model needs only to learn a binary classification of the two hypotheses and does not need to learn to find the most plausible explanation. They define αNLI task as follows:

- $\mathcal{O}_1$ and $\mathcal{O}_2$ are two observations at time $t_1 < t_2$.
- $h+$ is a positive (plausible) hypothesis and $h-$ is a negative (implausible) hypothesis.
- The αNLI task is to select the most plausible hypothesis from the $h+$ and $h-$.

We redefine αNLI as αNLI^*:

- $\mathcal{O}_1$ and $\mathcal{O}_2$ are two observations at time $t_1 < t_2$.
- h is a candidate hypothesis.
- The αNLI^* task is to test whether triple ($\mathcal{O}_1$, h, $\mathcal{O}_2$) is valid.

The approach of αNLI^* is similar to αNLI but different in the following ways:

- We ask the model to validate the triple instead of only choosing which hypothesis is more plausible amongst the given two;
- In αNLI, if the model did not choose a hypothesis, we still do not know whether it is valid, so we can only know that it is not plausible as the chosen one;
- In addition, αNLI^* can be feasible even when only one hypothesis or more hypotheses are given, which is not possible in αNLI.

5. Experiments and Discussions

Our experiments are designed to test the performance of several alternative models for binary classification tasks related to abductive reasoning using the L'ART dataset. Our extended dataset consists of 498,697 samples, of which 476,167 are used as the training set, 9,339 are used as the validation set, and 13,191 are used as the test set. The max-length in characters for observation 1, observation 2, and hypothesis in the training, validation, and test sets are shown in Table 2. Pre-trained transformer models have a built-in maximum token length for input, and ensuring that the maximum length does not exceed this limit helps avoid over-truncation issues. The number of samples in each class is identical in our binary classification setting.

Table 2
Max-length in characters for observation 1, observation 2, and hypothesis in the train, validation, and test sets.

	Train	Validation	Test
Observation 1 ($\mathcal{O}_1$)	71	71	71
Observation 2 ($\mathcal{O}_2$)	74	71	71
Hypothesis ($\mathcal{H}$)	184	147	150

In this experiment, it is crucial to use different training, validation, and test sets to ensure the evaluation of the selected transformer models is reliable and not influenced by overfitting to the training data. The training set is utilized for training the models, while the validation set is employed for tuning the hyperparameters and selecting the best model. Lastly, the test set assesses the final performance of the chosen model.

We chose specific train/test/validation split ratios to ensure that the models have adequate data for training while still maintaining ample samples for validation and testing. It is worth noting that deep learning approaches like these are inherently statistical in nature, and our validation/test sets, with approximately 10K samples each, provide a reliable reflection of the results.

We use several different transformer models in the experiment: the original BERT [30] (base and large version), and the top legal models (BERT-PLI [2], Legal BERT [3], BERTLaw[4] and NFSP version of Paralaw NetsParaLaw Nets [31]). We train each model on the input data, using the valid and invalid triples as the labels, and evaluate the performance of the models on the test set using accuracy. We also report the performance on our validation set. We run the experiment multiple times to ensure the reliability of the results, and we analyze the performance of the models on the test data to determine which model performs best on the binary classification tasks related to abductive reasoning. We also tested GPT-3 [32] zero-shot prediction and recorded the results on the test set of L'ART.

Table 3
Average performance of the models in three runs, sorted from high to low.

Model	Validation Accuracy	Test Accuracy
BERT Base	0.6202	0.6162
BERT-PLI	0.6115	0.6115
NFSP	0.5825	0.5808
Legal BERT	0.5756	0.5619
BERTLaw	0.5484	0.5371
BERT Large	0.5016	0.5000
GPT-3	-	0.4959

Table 3 shows the results of the experiment. Our first observation is that performance on the validation set and the test set are not significantly different, which is a good indication that the models aren't overfitting the training data or the hyperparameter tuning process. From those results, we can also observe that the original BERT Base model performs better on abductive reasoning than all state-of-the-art legal models and even the BERT Large model. This result is quite surprising because it means that the pre-trained legal models are not necessarily more effective than the original BERT Base model on abductive reasoning.

We believe this is because the legal models are trained on documents that are mostly directed toward legal reasoning, rather than abductive reasoning. In the legal domain, there are many documents that contain information that can help with legal reasoning, but there is not a lot of information in the legal domain that can help with abduction. This imbalance in the training data can lead to a bias in the training process that favors legal reasoning and harms the performance of the models on abductive reasoning.

The worst performance among finetuned models is BERT Large, which is also a surprise since it is usually reported to be a robust model on many NLP tasks. This result suggests that pretraining a model with a larger capacity with more data does not guarantee better performance. Adding commentary on GPT-3, which has the lowest performance on zero-shot

tasks, we included it for reference purposes, not for comparison, as the model has not been tuned for the specific domain and therefore cannot be fairly compared to the other models. Furthermore, a general comment can be made that the poor performance of these models indicates that abductive reasoning remains a challenging problem. The redefinition of the task within L'ART helps to make this issue more evident.

6. Conclusions

We have investigated the support for abductive reasoning provided by state-of-the-art (SOTA) transformer models within the legal field. To accomplish this, we first redefined the task of abductive reasoning and constructed a reliable dataset. Following this, we utilized the dataset to assess the performance of SOTA models in the legal sphere, as well as prominent large language models in Natural Language Processing (NLP). Our experimental results revealed that the SOTA models, including all legal-specific variants, do not necessarily outperform the original BERT Base model in abductive reasoning tasks. This outcome provides insight into current limitations when pretraining large language models for legal applications. The subpar performance of the original BERT Large and GPT models illustrates that simply increasing the size of the model and providing it with more data does not guarantee superior performance. Future directions for this research could include exploring alternative pretraining approaches specifically tailored to abductive reasoning tasks, developing novel architectures that focus on legal reasoning, and examining the relationship between model capacity and performance on abductive reasoning tasks. Additionally, further investigation into leveraging and integrating existing legal domain knowledge with the pretraining process may lead to more effective models capable of handling the unique challenges of legal reasoning tasks.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number, JP22H00543 and JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4. Francesca Toni and Kostas Stathis would like to thank the National Institute of Informatics, Tokyo, Japan, for supporting their visit to Japan that made this work possible. Francesca Toni also acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No.101020934, ADIX), as well as support from J.P. Morgan and the Royal Academy of Engineering, UK, under the Research Chairs and Senior Research Fellowships scheme.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[2] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, S. Ma, Bert-pli: Modeling paragraph-level interactions for legal case retrieval, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth

International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 2020, pp. 3501–3507. URL: https://doi.org/10.24963/ijcai.2020/484. doi:10.24963/ijcai.2020/484, main track.

[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).

[4] H.-T. Nguyen, H.-Y. T. Vuong, P. M. Nguyen, B. T. Dang, Q. M. Bui, S. T. Vu, C. M. Nguyen, V. Tran, K. Satoh, M. L. Nguyen, Jnlp team: Deep learning for legal processing in coliee 2020, arXiv preprint arXiv:2011.08071 (2020).

[5] V. Tran, M. Le Nguyen, S. Tojo, K. Satoh, Encoded summarization: summarizing documents into continuous vector space for legal case retrieval, Artificial Intelligence and Law 28 (2020) 441–467.

[6] M. Yoshioka, Y. Aoki, Y. Suzuki, Bert-based ensemble methods with data augmentation for legal textual entailment in coliee statute law task, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 2021, pp. 278–284.

[7] H.-T. Nguyen, M.-K. Phi, X.-B. Ngo, V. Tran, L.-M. Nguyen, M.-P. Tu, Attentive deep neural networks for legal document retrieval, Artificial Intelligence and Law (2022) 1–30.

[8] J. Rabelo, R. Goebel, M.-Y. Kim, Y. Kano, M. Yoshioka, K. Satoh, Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021, The Review of Socionetwork Strategies 16 (2022) 111–133.

[9] K. Abimbola, Abductive reasoning in law: Taxonomy and inference to the best explanation, CARDozo L. REv. 22 (2000) 1683.

[10] D. A. Schum, Species of abductive reasoning in fact investigation in law, in: The Dynamics of Judicial Proof, Springer, 2002, pp. 307–336.

[11] C. Bhagavatula, R. L. Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W. Yih, Y. Choi, Abductive commonsense reasoning, in: 8th International Conference on Learning Representations, ICLR 2020, OpenReview.net, Addis Ababa, Ethiopia, 2020. URL: https://openreview.net/forum?id=Byg1v1HKDB.

[12] P. M. Kien, H.-T. Nguyen, N. X. Bach, V. Tran, M. L. Nguyen, T. M. Phuong, Answering legal questions by learning neural attentive text representation, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 988–998. URL: https://aclanthology.org/2020.coling-main.86. doi:10.18653/v1/2020.coling-main.86.

[13] A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, H. Hajishirzi, XOR QA: Cross-lingual open-retrieval question answering, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 547–564. URL: https://aclanthology.org/2021.naacl-main.46. doi:10.18653/v1/2021.naacl-main.46.

[14] Y. Nie, S. Wang, M. Bansal, Revealing the importance of semantic retrieval for machine reading at scale, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2553–2566. URL: https://aclanthology.org/D19-1258. doi:10.18653/v1/D19-1258.

[15] J. Lee, C. Y. Yeung, Text retrieval for language learners: Graded vocabulary vs. open learner

model, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 798–804. URL: https://aclanthology.org/2021.ranlp-1.91.

[16] B. Krojer, V. Adlakha, V. Vineet, Y. Goyal, E. Ponti, S. Reddy, Image retrieval from contextual descriptions, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3426–3440. URL: https://aclanthology.org/2022.acl-long.241. doi:10.18653/v1/2022.acl-long.241.

[17] B. Y. Lin, D.-H. Lee, M. Shen, R. Moreno, X. Huang, P. Shiralkar, X. Ren, TriggerNER: Learning with entity triggers as explanations for named entity recognition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8503–8511. URL: https://aclanthology.org/2020.acl-main.752. doi:10.18653/v1/2020.acl-main.752.

[18] A. Ritter, S. Clark, Mausam, O. Etzioni, Named entity recognition in tweets: An experimental study, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 1524–1534. URL: https://aclanthology.org/D11-1141.

[19] J. Rabelo, M.-Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, K. Satoh, A summary of the COLIEE 2019 competition, in: New Frontiers in Artificial Intelligence, Springer International Publishing, Online, 2020, pp. 34–49.

[20] N. H. Thanh, B. M. Quan, C. Nguyen, T. Le, N. M. Phuong, D. T. Binh, V. T. H. Yen, T. Racharak, N. L. Minh, T. D. Vu, P. V. Anh, N. T. Son, H. T. Nguyen, B. Butr-indr, P. Vateekul, P. Boonkwan, A summary of the ALQAC 2021 competition, in: 2021 13th International Conference on Knowledge and Systems Engineering (KSE), IEEE, Bangkok, Thailand, 2021, pp. 1–5.

[21] C. Nguyen, M.-Q. Bui, D.-T. Do, N.-K. Le, D.-H. Nguyen, T.-T. Nguyen, H.-T. Nguyen, V. Tran, L.-M. Nguyen, N.-C. Le, T.-T. Le, M.-P. Nguyen, T.-B. Dang, T.-S. Nguyen, V.-A. Phan, T.-H.-Y. Vuong, M.-T. Nguyen, T. Le, T.-H. Nguyen, ALQAC 2022: A summary of the competition, in: 2022 14th International Conference on Knowledge and Systems Engineering (KSE), IEEE, Nha Trang, Vietnam, 2022, pp. 1–5. doi:10.1109/kse56063.2022.9953764.

[22] H.-T. Nguyen, M.-P. Nguyen, T.-H.-Y. Vuong, M.-Q. Bui, M.-C. Nguyen, T.-B. Dang, V. Tran, L.-M. Nguyen, K. Satoh, Transformer-based approaches for legal text processing, The Review of Socionetwork Strategies 16 (2022) 135–155. doi:10.1007/s12626-022-00102-2.

[23] A. Louis, G. Spanakis, A statutory article retrieval dataset in French, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6789–6803. URL: https://aclanthology.org/2022.acl-long.468. doi:10.18653/v1/2022.acl-long.468.

[24] Y. T.-H. Vuong, Q. M. Bui, H.-T. Nguyen, T.-T.-T. Nguyen, V. Tran, X.-H. Phan, K. Satoh, L.-M. Nguyen, SM-BERT-CR: a deep learning approach for case law retrieval with supporting model, Artificial Intelligence and Law (2022). doi:10.1007/s10506-022-09319-6.

[25] A. C. Kakas, R. A. Kowalski, F. Toni, Abductive logic programming, Journal of logic and computation 2 (1992) 719–770.

[26] P. Lertvittayakumjorn, I. Petej, Y. Gao, Y. Krishnamurthy, A. Van Der Gaag, R. Jago, K. Stathis, Supporting complaints investigation for nursing and midwifery regulatory

agencies, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, 2021, pp. 81–91.

[27] R. Jago, A. van der Gaag, K. Stathis, I. Petej, P. Lertvittayakumjorn, Y. Krishnamurthy, Y. Gao, J. C. Silva, M. Webster, A. Gallagher, et al., Use of artificial intelligence in regulatory decision-making, Journal of Nursing Regulation 12 (2021) 11–19.

[28] P. Clark, O. Tafjord, K. Richardson, Transformers as soft reasoners over language, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 2020, pp. 3882–3890. URL: https://doi.org/10.24963/ijcai.2020/537. doi:10.24963/ijcai.2020/537, main track.

[29] A. Gaskell, Y. Miao, F. Toni, L. Specia, Logically consistent adversarial attacks for soft theorem provers, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 2022, pp. 4129–4135. URL: https://doi.org/10.24963/ijcai.2022/573. doi:10.24963/ijcai.2022/573, main Track.

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[31] H.-T. Nguyen, P. M. Nguyen, T.-H.-Y. Vuong, Q. M. Bui, C. M. Nguyen, B. T. Dang, V. Tran, M. L. Nguyen, K. Satoh, Jnlp team: Deep learning approaches for legal processing tasks in coliee 2021, arXiv preprint arXiv:2106.13405 (2021).

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.