Evaluating Legal Accuracy of Neural Generators on the Generation of Criminal Court Dockets Description

Nicolas Garneau[†], Eve Gaumond[‡], Luc Lamontagne[†], and Pierre-Luc Déziel[‡]
Université Laval, Québec, Canada
Computer Science Department[†] and Faculty of Law[‡]
{nicolas.garneau,luc.lamontagne}@ift.ulaval.ca
eve.gaumond@observatoire-ia.ulaval.ca
pierre-luc.deziel@fd.ulaval.ca

Abstract

Docket files, also known as *plumitifs*, are legal text documents describing judicial cases. They are present in most jurisdictions and are meant to provide a window on legal systems. They contain information of a judicial case such as parties' identities, accusations' provisions, decisions, and pleas. However, this information is cryptic, using abbreviations, and making references to the criminal code. In this paper, we explore the use of neural text generators to improve the legal accuracy of the docket file verbalization regarding the accusations, decisions, and pleas sections. We introduce a legal accuracy evaluation scale used by jurists to manually assess the performance of three architectures with different levels of prior knowledge injection. We also study the correlation of our human evaluation methodology with automatic metrics.

1 Introduction

The *plumitif* [plymitif] is a legal registry providing short summaries of every judicial case heard by the courts in the province of Quebec, Canada. It is akin to what is known in English as court dockets, used in several other judicial systems. It provides information about the stakeholders involved in a case, the moment and the location where the various steps of the judicial process take place, and, in the case of the criminal *plumitif*, it also gives information about the offences, the pleas, and the verdicts. However, although this information is publicly available online, in reality, it is hardly accessible because of the format in which the *plumitif* is presented. It is written almost exclusively using abbreviations and makes numerous references to provisions in the Criminal code that are not defined anywhere (see Appendix A for an example). As a result, even experienced lawyers confess they sometimes have a hard time understanding the *plumitif* (Parada et al., 2020).

This lack of intelligibility is an issue (Tep et al., 2019; Parada et al., 2020; Beauchemin et al., 2020). Indeed, while the *plumitif* could serve many useful purposes, at the moment, it is not used to its full potential because of how hard it is to understand. For instance, the lack of intelligibility prevents self-represented litigants from using the *plumitif* to keep track of their cases. It also burdens the work of journalists using the *plumitif* to report on legal affairs. There are also instances of citizens who suffered prejudices because insurers misinterpreted their docket when they consulted it for background check purposes (Gaumond and Garneau, 2021). Therefore, tackling the issue of the understandability of Quebec's criminal *plumitif* is a worthy objective. It could promote access to justice, improve the transparency of the judicial system and prevent discrimination.

Beauchemin et al. developed a web application to tackle this issue. It works well to enhance the understandability of certain sections of the plumitif – the section about the parties involved in the case, for instance. However, the application, relying on a rule-based generator, lacks precision when it comes to generating a description of the charges. Indeed, it simply uses provisions' headings – as found in the Canadian Criminal code – to verbalize the charges. Hence, it would replace section 348 (1) of the Criminal code[1] with the following sentence: "Breaking and entering with intent, committing offence or breaking out." This does not take into account the nuances of section 348 (1), which provides for four different offences;

1. "breaks and enters a place *with intent* to commit an indictable offence therein"
2. "breaks and enters a place and *commits* an indictable offence therein"

[1] https://laws-lois.justice.gc.ca/eng/acts/c-46/section-348.html

3. “breaks out of a place *after*
    (a) committing an indictable offence therein
    (b) entering the place with intent to commit an indictable offence therein.

These four offences have different degrees of severity. For instance, a defendant breaking in somewhere with the intent of committing robbery could be remorseful and leave empty-handed the place he broke into. He would not be sanctioned as severely as another defendant who committed the robbery. Given the rule-based architecture Beauchemin et al. used, the only way for them to take more legal nuances into account would have been to “stitch” provision’s label with the corresponding paragraph and indent. Since a long stretch of text is known to be unintelligible (Gaumond and Garneau, 2021), this solution wasn’t suitable.

Instead, we propose to use neural architecture to generate descriptions of legal provisions that take legal nuances into account while being relatively concise. To that end, we trained neural text generators on Plum2Text (Garneau et al., 2021c), a Data-to-Text dataset, to solve this particular issue. However, neural architectures tend to hallucinate facts (Dušek et al., 2018) which raises a question regarding their usability to accomplish sensitive tasks such as ours. If these models were to hallucinate some information that does not appear in a docket – charges of which the defendant was not accused, for instance – they could not be used in a production setup.

In this paper, we propose a new legal accuracy evaluation scale used by jurists to manually assess the performance of the models we’ve trained. We analyze if they accurate enough to be used in sensitive tasks such as verbalizing the content of the *plumitif*. We thus provide a comparative study of three neural architectures and reflect on their performance from a legal standpoint. We also evaluate them using automatic evaluation metrics and study their correlation with a human evaluation.

2 Training Neural Networks on *Plum2Text*

In this section, we introduce the three models we will evaluate. First, we introduce the *Plum2Text* dataset designed to train language generators, and then, we proceed to present the selected neural architectures as well as their training procedure.

2.1 *Plum2Text*

For our experiments, we have access to *Plum2Text*, introduced by Garneau et al. (2021b). It is a data-to-text dataset designed to train neural architectures to generate short descriptions of court dockets. It is derived from the pairings between criminal court judgments (a long textual document) and their associated docket file. A training instance is depicted in Appendix B. *Plum2Text* contains 2,300 examples. The dataset is however heavily skewed towards common infractions such as “driving under the influence” (section 320.14 from the Canadian Criminal Code) and “assault and battery” (section 268). Our preliminary experiments showed that any neural text generator trained on *Plum2Text as-is* yielded models with poor generalization capabilities, often generating the most frequent offences. We thus undersampled *Plum2Text* so that every provision is represented by at most 5 examples. This undersampling yields a dataset of 1,602 examples that we split randomly into a train, valid, and test sets which contain 931, 247, and 424 examples respectively. To better assess the generalization capabilities of the generators, we identified 9 provisions in the test set that are neither in the train or valid sets. The provisions are listed in Table 3, and we provide more details on the results of these specific examples in the evaluation Section 3.

2.2 Neural Text Generators

Neural architectures have proven to be very effective at generating text in a wide variety of tasks. Long-Short Term memory networks using attention achieved impressive performance on automatic machine translation (Bahdanau et al., 2015). GPT, the Generative Pretraining architecture, based on the Transformer architecture (Vaswani et al., 2017), pushed automatic textual generation to a whole new level not only for machine translation but also for text summarization and data-to-text generation (Radford et al., 2018, 2019; Brown et al., 2020). The performance of GPT is largely due to prior knowledge injection where fine-tuning on a downstream task requires less training data. Indeed, this model has been pre-trained on a large corpus before being trained on the target task. Prior knowledge injection is highly effective, especially when the prior is closer to the downstream task in terms of semantics and lexical field (Howard and Ruder, 2018). We thus consider three models, each with their respective degree of prior knowledge injection. The first one is the model proposed by Bahdanau et al. (2015) trained from scratch on *Plum2Text* (*no prior*) using the same procedure as Wiseman et al. (2017). We then selected a French pre-trained model based on the Transformer network, BARThez (Kamal Eddine et al., 2021)[2] (*language prior*). The last model we consider is a fine-tuned version of BARThez on a legal corpus, *Criminel*BART (Garneau et al., 2021a) (*language and domain prior*).

In order to conduct our experiments, we used the fairseq library[3] which provides implementations for the three models introduced in Section 2. For the three models, we used at most 1024 tokens (which resulted in batch sizes of 10 examples on average), the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0005, a cosine learning rate scheduler, and a dropout of 0.1. The LSTM model converged after 15 epochs. BARThez converged after 2 epochs, and *Criminel*BART after only one epoch. We trained all models on a GeForce 2080 using a personal desktop. Each training takes less than an hour to run. The LSTM has around 3M parameters, while BARThez and *Criminel*BART have both 139M parameters. For the generation of legal descriptions, we used beam search, with a beam size of 6. We post-process the generations by detokenizing the sentences to increase the readability. At inference time, the LSTM model takes on average 0.1 seconds per generation, BART 1 second and *Criminel*BART 0.5 second. As a matter of reproducibility, we provide all the generations of the models [4]. We provide generation examples in Appendix C.

3 Evaluation

As explained in section 1, the goal of this paper is to determine if neural architectures are accurate enough to be used in sensitive tasks such as verbalizing the content of the *plumitif*. Putting it another way, we want to see if some of the evaluated models could be used in a production setup. We also aim to characterize the strengths and weaknesses of neural generators in the field of law. We first introduce our methodology, discuss our expectations and finally analyze the results.

3.1 Methodology

We first introduce new human evaluation guidelines motivated by the underlying task of measuring the legal accuracy of the models. We then analyze the performance of the models using several automatic evaluation metrics. Finally, we analyze the correlation between automatic and human evaluation in order to ground one or several metrics in the context of automatic model selection.

3.1.1 Human Evaluation

Generating descriptions of criminal court dockets – which we trained our neural-text-generators for – is a rather sensitive task. Inaccurate generations run contrary to the very objective we pursue and could have real consequences. For example, imagine the potential harms resulting from a docket description that says that a defendant is guilty of a charge, while he was actually acquitted; that he was accused of possessing child pornography while he was actually accused of possession of cannabis; or that he pleaded guilty while it was not the case. This is why we deemed it essential to assess the quality of the generations not only from a technical standpoint but also from a legal standpoint. Following the arguments of van der Lee et al. (2019), we answer several questions regarding the experimental setup and the choices we made.

Selected models. We selected all three models trained on *Plum2Text*, allowing us to evaluate the improvement of prior knowledge injection in the field of legal text generation.

Number of outputs. From the selected test set containing 232 instances, we carefully selected instances yielding a diverse sample for the annotators to evaluate. This resulted in 89 instances, 64 with one table value and 17 with two table values. We manually created 8 instances containing three table values (*i.e.*, provision, decision, and pleading) since no instance contain the three different types of values in the original test set. Each instance is associated with three output generations, yielding a total of 267 outputs to evaluate.

Input selection. Following the recommendations of van Miltenburg et al. (2021), we select specific kinds of inputs and analyze their corresponding outputs. Hence, we begin by presenting to the annotator simple inputs containing only one

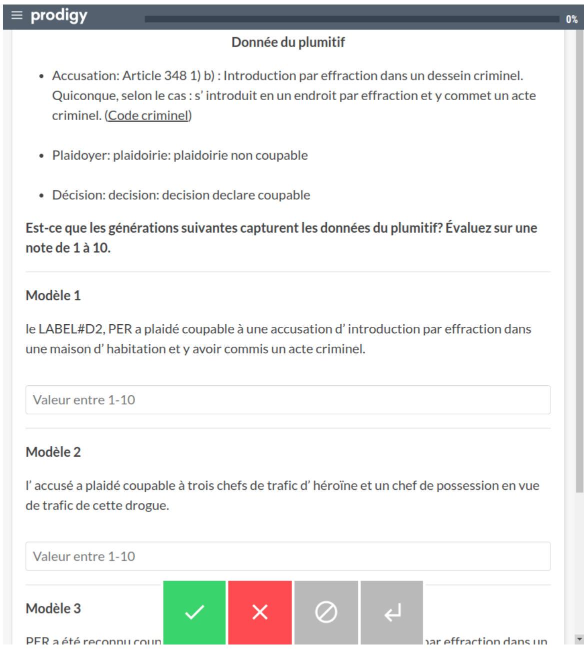
[2] BARThez is the French version of BART(Lewis et al., 2020), which showed interesting performance across several datasets for the task of data-to-text generation (Gehrmann et al., 2021).

[3] https://github.com/pytorch/fairseq

[4] https://drive.google.com/file/d/1JT6j9X47bjs54rEXkdL_OBELcmH0ZN6a/view?usp=sharing

table value. We then gradually increase the complexity of the inputs going up to three table values, which represent a whole *plumitif*'s line (charge, pleading, verdict, except for the sentence). This procedure allows the annotators to become familiar with the annotation interface, the dataset, and the task. We can also analyze the performance of the models given the inputs' increasing complexity.

Presentation and interface. We used the Prodigy annotation tool (Montani and Honnibal, 2018) and customized it for our need to present the *plumitif*'s input data and the three models' outputs. For each instance, the outputs are randomly ordered. The annotators are asked to score each of the three models' generation independently. This way of characterizing generations' relevance simultaneously has proven to be highly efficient in a model selection setup (Novikova et al., 2018). The evaluation interface is illustrated in Figure 1.



Donnée du plumitif

- Accusation: Article 348 1) b) : Introduction par effraction dans un dessein criminel. Quiconque, selon le cas : s' introduit en un endroit par effraction et y commet un acte criminel. (Code criminel)
- Plaidoyer: plaidoirie: plaidoirie non coupable
- Décision: decision: decision declare coupable

Est-ce que les générations suivantes capturent les données du plumitif? Évaluez sur une note de 1 à 10.

Modèle 1

le LABEL#D2, PER a plaidé coupable à une accusation d' introduction par effraction dans une maison d' habitation et y avoir commis un acte criminel.

Valeur entre 1-10

Modèle 2

l' accusé a plaidé coupable à trois chefs de trafic d' héroïne et un chef de possession en vue de trafic de cette drogue.

Valeur entre 1-10

Modèle 3

PER a été reconnu coup

par effraction dans un

Figure 1: The Prodigy annotation interface used by the annotators to semantically evaluate the generation of the three neural architectures.

Annotators. We selected three annotators with legal knowledge to evaluate the generations. The first annotator is a coauthor of this paper. She holds a bachelor's degree in law. Since she mainly works on the legal aspects of the project, she had not seen any of the generations nor any of the models before conducting the evaluations. She advised the principal investigator in the drafting of the evaluation guidelines and went through the evaluation process before the two other annotators to ensure that the guidelines were sufficiently clear for people trained in law. Her results should be read with all of that in mind. The other two annotators are second-year law students at the Faculty of Law. They were introduced to the context, the task, and the annotation interface in a meeting with the principal investigator and the first annotator. Another meeting was also held after a pilot evaluation. During this meeting, annotators 2 and 3 – who by then had evaluated 5 instances (i.e. 15 generations) – received feedback and advice on what phenomena they should be careful for. Annotators are paid at an hourly rate of 17 CAD/hour. Annotators were asked to spend at most 5 minutes per instance. It took a total of 8 hours for each annotator to complete the evaluation, including the training, the pilot and reading the evaluation guidelines.

After the annotators completed the evaluation, we gathered their comments on the difficulty of the task and if they encountered ambiguous cases. It turned out that the provisions' texts can be ambiguous since they may contain some disjunction in regards to the committed offence. Take for example provision 320.14 (1) a), "Operation while impaired", which states that "Everyone commits an offence who operates a conveyance while the person's ability to operate it is impaired to any degree by *alcohol or a drug or by a combination of alcohol and a drug*;". For this provision, models always generated a description only regarding the "degree of alcohol", omitting the drug aspect of the offence. However, one would need to look at the judgment file (if any) to validate if the defendant operated the conveyance impaired by alcohol or a drug or a combination of both. In such cases, annotators were unsure if the generation contained hallucinated/omitted facts, since the generation was not totally supported by the docket file's data. We can thus conclude that given a high agreement score and several ambiguous cases suggest that our instruction were clear for the annotators and the legal accuracy scale was easy to use. We provide in-depth details of the evaluation setup in our Human Evaluation Datasheet (Shimorina and Belz, 2021) in Appendix E.

Legal Accuracy Scale. Given that our aim with this paper is to determine whether or not neural-text generators are sufficiently accurate to be used in a production setup, we needed a definition of the notion of legal accuracy for our particular context

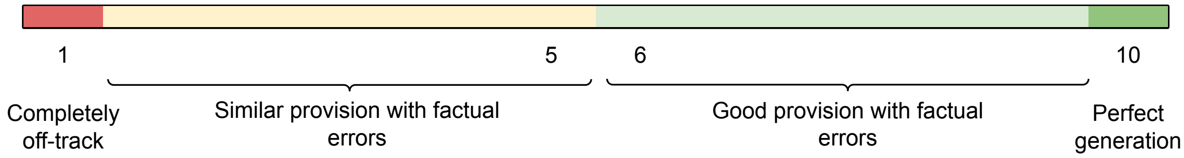

1
5
6
10
Completely off-track
Similar provision with factual errors
Good provision with factual errors
Perfect generation

Figure 2: The legal accuracy scale used by the human annotators. The annotators first decide where, between the four regions (completely off-track, similar provision, good provision, and perfection generation) the actual generation sits. Then they remove points for every hallucination and/or omissions encountered.

as well as an assessment tool to measure it. We thus define "legal accuracy" as a metric that measures the congruence between the docket's description that our models generate and the input data (Criminal code provision, plea, and verdict) the model aims to describe. To this end, we designed a Likert scale ranging from one to ten (Likert, 1932). A generation that scores a ten is highly accurate. However, a generation receiving a score of 1 misses the mark as it does not match the input data at all. To determine the legal accuracy score that a generation should receive, the annotator has to follow a two-step process. First, they decide if the docket description is 1) accurate, 2) thematically relevant, or 3) off-track. The legal accuracy scale is split into three regions;

1. **6 - 10 – Accurate.** If the generation refers to the good provision, it is considered accurate and will score be between 6 and 10.

2. **2 - 5 – Thematically Relevant.** If the generation is "on theme" with the input data, the score will be between 2 and 5. A thematically relevant description is related to the right provision, but not perfectly on point (possession of drugs vs possession of weapons; sexual exploitation vs child pornography; breaking in with the intent of committing a crime vs breaking in and committing a crime).

3. **1 – Off-Track.** If the generation is about "Mischief" while the input was about "Drug trafficking", we ask the annotator to assign a score of 1 since it is completely off-track.

Once the annotators have chosen the bracket where the generation belongs (1 ; 2 - 5 ; 6 - 10)[5], they can start moving on to the second step: looking for factual errors. We identify three types;

1. Hallucinations: facts that the model generates even though it does not appear in the input data. There are various kinds of hallucinations: the model generates a charge, verdict, or plea that does not appear in the *plumitif*, provides some factual details about the perpetration of the infraction that should not appear in the *plumitif* (e.g. the defendant did traffic *heroin* unlawfully). One point should be removed per hallucination.

2. Omissions: occurs when facts are in the input data but end up not being generated by the model. One element that is quite often omitted is the provision number. The absence of the plea or the verdict is also considered an omission when it was available in the input data. One point should be removed per omission.

3. Confusions: factual mistakes characterized by the mismatch of the input data and the content of the generation. For example, the input data says that the defendant pleaded guilty while he appears to have pleaded not guilty in the generation, or the court orders a stay of procedures in the *plumitif*, and the defendant is found guilty in the generation. In these cases, two points should be removed: one for hallucinating a fact, and one for omitting a fact.

No matter how many factual errors there are, they can't make a generation downgrade to the lower bracket. So, a thematically relevant generation can't have less than 2 points, and a generation

[5] At first, we split the scale into four regions: 1; 2-5; 6-9; 10. However, the annotators tend to naturally split it into three regions since they can not directly attribute 10 points to a given generation whereas they can directly attribute 1 point to an irrelevant generation. They first need to see if the generation is accurate, on theme, or irrelevant, before proceeding to the second step.

that gets the provision right can't have less than 6 points. Finally, a provision from the 6-10 points bracket which is exempt from factual errors, gets 10 points, which is the highest possible mark on our legal accuracy scale. To summarize this process, Figure 2 provides a conceptual illustration of the Likert scale.

3.1.2 Automatic Evaluation

For the automatic evaluation of the neural models, we use the same set of commonly used metrics as in the GEM benchmark suite (Gehrmann et al., 2021). We can differentiate the metrics according to two features: those using surface tokens or vector representations, and those using the reference and/or the table values. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) use surface tokens and a reference. BertScore (Zhang* et al., 2020) (using the underlying multilingual version of BERT (Devlin et al., 2019)[6]) uses vector representation and a reference. We also consider two metrics using vector representation and the table values recently introduced by Dušek and Kasner (2020) and Garneau and Lamontagne (2021), which we dubbed respectively NLI and RANK in this paper.

NLI uses natural language inference to check if a given hypothesis entails or contradicts table values. According to the methodology described by Dušek and Kasner (2020), we created three templates needed for the computation of the NLI metric. These three templates are each associated to a specific type of value, which are the accusation, the plea, and the verdict that take as input a subject and an object[7];

- `<subject>` is accused of `<object>` .
- `<subject>` pleaded `<object>` .
- `<subject>` is declared `<object>` .

Given the table values, we fill in these templates and perform natural language inference w.r.t. the hypothesis under test. We used the pre-trained CamemBERT (Martin et al., 2020) base NLI model in our experiments.

RANK uses a ranking model coupled with the mean average precision to assess the ability of a given hypothesis to retrieve its corresponding table values. According to the methodology described by Garneau and Lamontagne (2021), we trained the multilingual version of BERT using the *plum2text* dataset on the semantic textual similarity task. This yielded a model able to rank table values w.r.t. the hypothesis under test, as required by RANK.

A reference-less metric can be interesting in cases where we do not have access to manually curated pairs of table–reference or in a production setting where references are simply nonexistent. As Zhang* et al. (2020) suggest, metrics using embeddings instead of surface tokens showed better correlation with human evaluation in several settings, a phenomenon we wish to confirm in our setup.

3.1.3 Grounding Metrics

Finally, we wish to ground automatic evaluation metrics w.r.t the legal accuracy scores to speed up the model selection process, which would be highly desirable in a concrete application setup (Belz and Reiter, 2006; van der Lee et al., 2019). To this end, we compute the Spearman correlation of the human evaluation scores with every automatic metric introduced in the previous section.

3.2 Expectations

According to the goal and the human and automatic evaluation methodologies previously introduced, we have the following expectations regarding the experiments;

1. We expect that models containing more prior knowledge on the downstream task will perform better and may have better generalization capabilities, as exposed by (Peters et al., 2018; Devlin et al., 2019; Raffel et al., 2020). This supports our approach of using the models mentioned previously with three different levels of prior knowledge.

2. We do not expect high correlation scores between human evaluation and metrics based on word overlap (BLEU, ROUGE, METEOR) (Belz and Reiter, 2006; Novikova et al., 2017). However, we expect better correlation scores with metrics that use vector representations (BertScore) and use the input table for their computation (NLI, RANK) (Zhang* et al., 2020).

3. We expect that the increasing complexity of the input (i.e. adding the verdict and the plea

[6] We did not use BLEURT (Sellam et al., 2020) since it has been trained on an English corpus.

[7] The subject is always the same, being "*The defendant*"

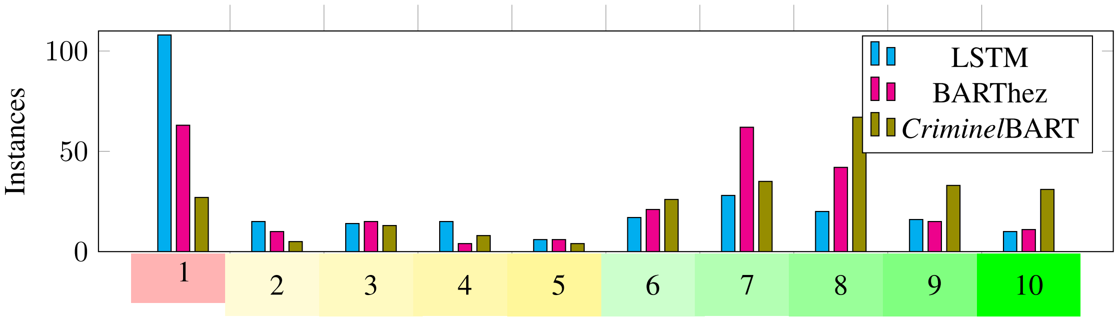


Figure 3: Results of the human evaluation according to the legal accuracy scale. We present the results of the vanilla LSTM (no prior), BARThez (language prior), and *Criminel*BART (language and domain prior).

	LSTM	**BARThez**	***Criminel*BART**
Ann. 1	4.4±2.8	5.2±2.9	6.3±2.6
Ann. 2	3.7±3.2	5.2±3.0	6.8±2.8
Ann. 3	3.6±3.3	5.4±3.2	7.0±2.8
Avg.	3.9±2.9	5.3±2.9	6.7±2.6
ρ	0.76	0.85	0.84

Table 1: Average score and standard deviation per annotator and the overall score for each model. We also provide the annotator agreement ρ per model. The overall agreement is 0.84.

as input) should not impact the models' performance dramatically since the range of values of this type of data is limited (e.g. up to 10 different verdicts and two different pleas).

3.3 Results

In this section, we first analyze how automatic metrics correlate with human judgment. We then study the benefit of language and domain prior knowledge injection, both on seen and unseen distributions of the data. We also diagnose the learning dynamics of the neural architecture w.r.t the increasing complexity of the input.

3.3.1 Prior knowledge

Results of the human evaluation on the 267 outputs are displayed in Figure 3 using the legal accuracy scale. We can see that the LSTM model has difficulty finding itself on the right side of the scale, having more than 100 irrelevant generations and achieving an overall score of 3.9. BARThez, containing a substantial language prior, does perform much better than the vanilla LSTM, achieving an average score of 5.3 mostly due to its 60 irrelevant generations. Its generations are mostly spread on the far left, and middle right of the scale. *Criminel*BART achieves the best performance with an overall score of 6.7, having most of its generation containing the "good provision". From these results, and w.r.t. the legal accuracy scale, this tells us that on average, *Criminel*BART will be on theme with possibly 2-3 hallucinations/omissions. This observation validates our first expectation regarding the contribution of prior knowledge injection.

We also provide the breakdown of the scores by annotator in Table 1. Annotator 1 provided scores on a narrow scale, ranging from 4.4 to 6.3 on average, whereas Annotator 2 and 3 used a wider scale with scores ranging from 3.6 to 7.0. Since we have multiple annotators and an ordinal scale, we used Krippendorff's alpha coefficient (Krippendorff, 2004) to measure inter-annotator agreement. We obtained a correlation coefficient ρ of 0.84 across all models. This high correlation coefficient suggests that either the evaluation task was easy and/or the evaluation guidelines were clear and easily understood by the annotators. Looking at the agreement model-wise, we obtained a ρ coefficient of 0.85 and 0.84 for the BARThez and *Criminel*BART evaluations, respectively. For the LSTM model, we obtained a ρ coefficient of 0.74. It seems like the annotators tend to disagree when the generations are worse, probably misclassifying a generations as being "on theme" (2 - 5) or "irrelevant" (1).

In Table 2, we present the results of the experiments using automatic evaluation metrics. One of our expectations was that the more prior knowledge a model has, the better it will perform. While *Criminel*BART is the best model across all metrics, it is interesting to see however that, according to the metrics using references, BARThez performs worse than the vanilla LSTM. On the other hand, by looking at the metrics using the table values,

	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE-L**	**METEOR**	**BERTScore**	**NLI**	**RANK**
LSTM	0.38	0.28	0.23	0.20	0.33	0.20	0.75	0.28	0.21
BARThez	0.32	0.24	0.19	0.16	0.34	0.21	0.74	**0.34**	0.38
*Criminel*BART	**0.51**	**0.42**	**0.36**	**0.32**	**0.44**	**0.28**	**0.78**	**0.34**	**0.43**

Table 2: Automatic evaluation results of the three models using token-based metrics (BLEU, ROUGE, and METEOR) and embedding-based metrics (BERTScore, NLI, and RANK).

Provision	**LSTM**	**BARThez**	***Criminel*BART**
445.1 (1) a)	1.0	1.0	1.0
150	2.3	5.0	4.6
83.181	1.0	1.0	1.0
241	1.0	2.7	2.0
467.12	1.0	1.0	8.7
810.2	1.0	1.0	1.0
172	1.0	1.0	1.33
320.14	1.0	6.3	7.3

Table 3: Analysis of the generalization capabilities of the models on unseen provisions. We provide details on the provisions in Appendix D.

BARThez seems to be substantially better than the LSTM model. From these results, it is not clear if the language prior was truly beneficial in our setup. However, the domain prior improves substantially the performance of the generations.

Finally, we analyze the generalization capabilities of the models on *unseen provisions i.e.*, provisions that were included neither in the training nor in the validation sets. We identified 8 unseen provisions, listed in Table 3. The results show that all the models struggle to fully generalize to unseen provisions. We can see that the LSTM can not generalize to unseen provision, which is expected. An interesting fact is that even if BARThez does not have any domain prior, it generalizes as well as *Criminel*BART except for one provision, 467.12, which corresponds to "Commission of offence for criminal organization". While BARThez and *Criminel*BART achieve a decent performance on provision 320.14 (Operation while impaired), it is more of a training set artifact since provision 253 that has been repealed in 2018 also corresponding to "Operation while impaired" is present in the training set. In a similar vein, the repealed provision 150 corresponding to "having illegally in his possession for sale magazines that are obscene" is similar to several many other charges of a sexual nature in the training set (e.g. 163, "Obscene materials") explaining why every model are "on-theme" for this provision.

3.3.2 Correlations Between Human and Automatic Evaluations

In every case, results show a positive correlation between human evaluation and automatic metrics. Word overlap metrics (BLEU-x, ROUGE-L, and METEOR) tend to show decreasing correlation scores as the model produces better generations; going from 0.4 with the LSTM to 0.2 with *Criminel*BART. BERTScore, an embedding-based metric, presents a high correlation score with the LSTM model. However, regarding BARThez and *Criminel*BART, correlation scores drop as low as 0.12. NLI provides consistent correlation scores of 0.35 on average, regardless of the model. RANK offers the highest correlation scores w.r.t. the models, reaching 0.81 with the LSTM model, 0.62 with BARThez, and 0.40 with *Criminel*BART. We suppose that these high correlation scores are tied to the nature of the last two metrics; they are using the input values as a way to assess the relevance of the generation, thus measuring its factual accuracy. On the other hand, overlap-based metrics and BERTscore only use the target reference which may not capture the factual accuracy one may be looking for. In light of these results, we deem it possible to use one or several metrics grounded with the proposed human evaluation to select the best-performing model for futur works.

3.3.3 Increasing Complexity of the Input

To better understand the learning dynamics of the neural architectures, we analyze their performance w.r.t. to the increasing complexity of the input i.e., going from one to three table values. More precisely, we want to study how models are able to combine semi-structured information that has been "linearized" as in Wiseman et al. (2017). Looking at Table 5, we can see that the performance of the LSTM model rapidly decreases as we add values in the input, going from 4.8 with one value, to 2.0

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore	NLI	RANK
LSTM	0.39	0.41	0.37	0.36	0.45	0.39	0.39	0.32	**0.81**
BARThez	0.25	0.31	0.31	0.28	0.38	0.40	0.12	0.35	**0.62**
*Criminel*BART	0.20	0.21	0.20	0.19	0.26	0.28	0.25	0.30	**0.40**

Table 4: Spearman correlation scores of automatic metrics with human evaluation. All scores have a p-value < 0.05 except for the pairs BARThez–BERTScore and *Criminel*BART–BLEU-x, which exhibit the lowest correlations. We highlighted in bold "row-wise" highest correlations, showing that RANK has capabilities to select the best model.

	LSTM	BARThez	*Criminel*BART
1 Value	4.8±2.9	5.7±2.8	6.8±2.7
2 Values	2.0±1.0	4.6±2.9	7.3±1.9
3 Values	1.0±0.0	3.9±2.8	4.5±1.7

Table 5: Analysis of the increasing complexity of the input by models, going from one to three table values.

with two, and 1.0 with three. Unfortunately, the model is not able to generate relevant descriptions as the complexity increases. We can also see a slight decrease in performance with the BARThez model going from 5.7 to 4.6 and 3.9 for one, two, and three input values respectively. *Criminel*BART, on the other hand, did maintain relevant generations with two input values with an average score of 7.3. However, its performance decreases with three table values, dropping at 4.5 on average. This analysis suggests that generating the complete line of a docket file (*i.e.*, accusation, decision, and pleading) is not properly handled by the neural architectures and that more training data would be beneficial. This observation invalidates our expectation that adding the verdict and the plea does not impact the models' performance.

4 Conclusion

In this paper, we evaluated the performance of three neural architectures, both automatically and manually, on the Data2Text task of docket files description generation. We proposed a new 10-point Likert scale to assess the legal accuracy of these architectures. We studied the correlation of automatic metrics with our human evaluation methodology and found out that the RANK metric can be used for automatic model selection. We release the generations of all three models as well as their associated automatic and human (anonymous) evaluation scores for a matter of reproducibility and for the research community's benefit. Unsurprisingly, *Criminel*BART is the best performing model due to its prior knowledge of the legal field. On average, it generates descriptions containing the good provision and better handles the increasing complexity of the input. However, its hallucination and omission rates suggest the need for improvements in this regard to obtain acceptable legal accuracy. Future works will look at better ways to condition this model to improve its legal accuracy using hard constraints (Meister et al., 2020) and post-edition (Mallinson et al., 2020). However, we believe that these models will require a human validation to be used in production, due to their inherent probabilistic nature and the sensitive legal field. We further discuss this matter, as well as the ethical considerations of having such a model in production in the following Section 5; **Broader Impacts – Law and Ethics**.

Acknowledgements

We thank the reviewers for their insightful comments as well as David Beauchemin, Ulysse Côté-Allard, Mathieu Godbout, Jean-Samuel Leboeuf, and Frédérik Paradis from the *Groupe de recherche en apprentissage automatique de Laval* who peer-reviewed our manuscript. This research was enabled in part by the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Social Sciences and Humanities Research Council (SSHR). Also, this research was made possible with the help and contribution of our partner, *La Société Québecoise d'Information Juridique*. Lastly, we wish to thank the annotators who evaluated the generations, Andréa Lampron and Noémie Lamothe.

5 Broader Impacts – Law and Ethics

As discussed in the introductory part, Quebec's plumitif is hard to understand. This well-documented issue (Parada et al., 2020; Tep et al., 2019; Prom Tep et al., 2020; Beauchemin et al., 2020) hinders access to justice, causes prejudices to people subject to background checks and contributes to a certain opacity of the judicial system (Gaumond and Garneau, 2021). Beauchemin et al. developed a web application to address this issue, but the performance of their rule-based text-generator is not satisfactory w.r.t the description of the charges. We thought that an alternative architecture, based on neural networks, could improve the charges' description. However, we were uncertain about the legal accuracy of neural-text generators knowing their propensity to hallucinate facts (Dušek et al., 2018). Therefore, we designed an evaluation method to assess the legal accuracy of three neural models generating descriptions of criminal charges. This process leads to the conclusion that *Criminel*BART is – with an average score of 6.7/10 – the best model to generate descriptions of criminal charges appearing in Quebec's plumitif. In the next sections, we reflect on what is required, in terms of legal accuracy.

5.1 What Is Considered Accurate Enough?

AI technologies used in the legal system ought to reach a high level of accuracy. This is obvious when we think about predictive tools informing judges' decisions (Surden, 2020) such as COMPAS, the infamous recidivism prediction algorithm (Dressel and Farid, 2018). But it should be equally clear that accuracy is crucial for AI systems used to disseminate judicial information. The intended purpose of an AI system determines the level of accuracy it should meet. *Criminel*BART aims at reducing the number of errors people make when they access the plumitif. It's a purpose that commands a high degree of accuracy. Indeed, if its generations are inaccurate, *Criminel*BART is both useless and dangerous. Useless because it goes against the very purposes it tries to achieve; and dangerous because providing inaccurate information about people's criminal history could lead to harm such as discrimination.

5.2 Is *Criminel*BART accurate enough?

We voluntarily chose not to pinpoint where the legal accuracy threshold falls; we don't want to say that a score of 9.5 on our scale means that a model is ready for production. Determining if a model is ready to move to production is contextual. A specific risk assessment should be done to make such a determination. In this case, the conclusion is that *Criminel*BART isn't accurate enough. With an average of two or three factual mistakes per generation – and even more inaccuracies when it comes to unseen provisions – *Criminel*BART is not ready to be used in a production setup. The example below provides an illustration:

- On REDACTED DATE, at REDACTED PLACE, the defendant broke and entered a dwelling-house with the intention to commit an offence therein, thus committing the indictable offence provided at section 348(1)b)d) of the Criminal Code.

There are four problems with this generation. First, there is one offence – sexual assault – that doesn't appear in the generation even though it was inputted into the model. Second, the generation says that the break-in happened in a dwelling-house while no such information was input into the model. This hallucination could be consequential since break-ins in dwelling-houses are considered more serious, and receive longer sentences. Third, the date and location of the offence are also hallucinated. Finally, the provision number should have been 348(1)b) instead of 348(1)b)d).

5.3 How to Increase Legal Accuracy?

Given the high degree of accuracy required for our purposes, it is not clear that neural text-generators will ever be accurate enough to be used without human oversight. Combining computers and humans' strength to increase *Criminel*BART's accuracy might be the way forward. Since writing descriptions of the Criminal code's provisions is a tedious task unlikely to be undertaken by humans, *Criminel*BART could generate drafts that court clerks would post-edit for accuracy. However, clerks are already tied-up. To ensure the adoption of the technology, this new post-edition task shouldn't feel burdensome to them. Players in the field make the success of legal innovations. It's important to make sure that their opinion is heard and considered and that they see the innovation as presenting some advantages for them (Benyekhlef et al., 2016).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

David Beauchemin, Nicolas Garneau, Eve Gaumond, Pierre-Luc Déziel, Richard Khoury, and Luc Lamontagne. 2020. Generating intelligible plumitifs descriptions: Use case application with ethical considerations. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 15–21, Dublin, Ireland. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Karim Benyekhlef, Jane Bailey, Jacquelyn Burkell, and Fabien Gelinas, editors. 2016. *eAccess to Justice*. Law, Technology and Media. University of Ottawa Press, Ottawa, ON, Canada.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1).

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2021a. Criminelbart: A french canadian legal language model specialized in criminal law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 256–257, New York, NY, USA. Association for Computing Machinery.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2021b. Plum2text: A french plumitifs-descriptions data-to-text dataset for natural language generation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 200–204, New York, NY, USA. Association for Computing Machinery.

Nicolas Garneau, Eve Gaumond, Luc Lamontagne, and Pierre-Luc Déziel. 2021c. Plum2text: A french plumitifs–descriptions data-to-text dataset for natural language generation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, Sao Paulo, Brazil. International Association for Artificial Intelligence and Law.

Nicolas Garneau and Luc Lamontagne. 2021. Trainable ranking models to evaluate the semantic accuracy of data-to-text neural generator. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eve Gaumond and Nicolas Garneau. 2021. 5q ia clartÉ plumcr qc : Cinq questions permettant d'appréhender l'usage d'intelligence artificielle pour accroître la clarté du plumitif criminel québécois. In Lex Electronica, editor, *La justice dans tous ses états*, pages 216–248. Montréal.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Rensis Likert. 1932. *A technique for the measurement of attitudes*. Archives of psychology ; no. 140. [s.n.], New York.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: Flexible text editing through tagging and insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alexandra Parada, Sandrine Prom Tep, Florence Millerand, Pierre Noreau, and Anne-Marie Santorineos. 2020. Digital Court Records : a Diversity of Uses. *IJR*, 9.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sandrine Prom Tep, Florence Millerand, Alexandra Bahary-Dionne, Sarah Bardaxoglou, and Noreau Pierre. 2020. Le "plumitif accessible" : les enjeux liés à l'accès aux registres informatisés en ligne. In Yvon Blais, editor, *22 chantiers pour l'accès au droit et à la justice*, pages 43–66. Montréal.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Anastasia Shimorina and Anya Belz. 2021. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in nlp. *ArXiv*, abs/2103.09710.

Harry Surden. 2020. Ethics of AI in law.

Sandrine Prom Tep, Florence Millerand, Alexandra Parada, Alexandra Bahary, Pierre Noreau, and Anne-Marie Santorineos. 2019. Legal Information in Digital Form: the Challenge of Accessing Computerized Court Records. *IJR*, 8.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

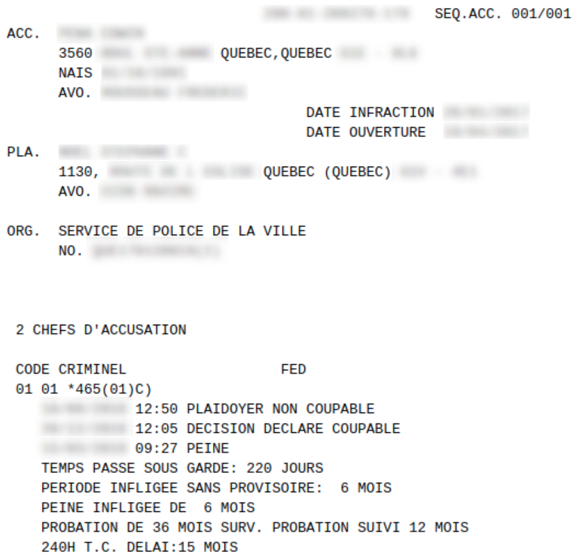
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Example Docket File

An example docket file is depicted in Figure 4. The accusation section, starting at the middle of the document, contains provisions' number, paragraph and indent (*465(01)c) in the figure) as well as the associated decision and plea (PLAIDOYER NON COUPABLE and DECISION DECLARE COUPABLE).



```
                                   [redacted]   SEQ.ACC. 001/001
 ACC.  [redacted]
       3560 [redacted] QUEBEC,QUEBEC [redacted]
       NAIS [redacted]
       AVO. [redacted]
                                DATE INFRACTION [redacted]
                                DATE OUVERTURE  [redacted]
 PLA.  [redacted]
       1130, [redacted] QUEBEC (QUEBEC) [redacted]
       AVO. [redacted]

 ORG.  SERVICE DE POLICE DE LA VILLE
       NO. [redacted]


  2 CHEFS D'ACCUSATION

  CODE CRIMINEL                  FED
  01 01 *465(01)C)
     [redacted] 12:50 PLAIDOYER NON COUPABLE
     [redacted] 12:05 DECISION DECLARE COUPABLE
     [redacted] 09:27 PEINE
     TEMPS PASSE SOUS GARDE: 220 JOURS
     PERIODE INFLIGEE SANS PROVISOIRE:  6 MOIS
     PEINE INFLIGEE DE  6 MOIS
     PROBATION DE 36 MOIS SURV. PROBATION SUIVI 12 MOIS
     240H T.C. DELAI:15 MOIS
```

Figure 4: A *Plumitif* document presenting the defendant and plaintiff personal information along with charges and associated pleas, decisions and penalty. Some regions have been blurred for privacy concerns.

B *Plum2Text* Training Instance

A training instance from the *Plum2Text* training dataset pictured in Figure 5.

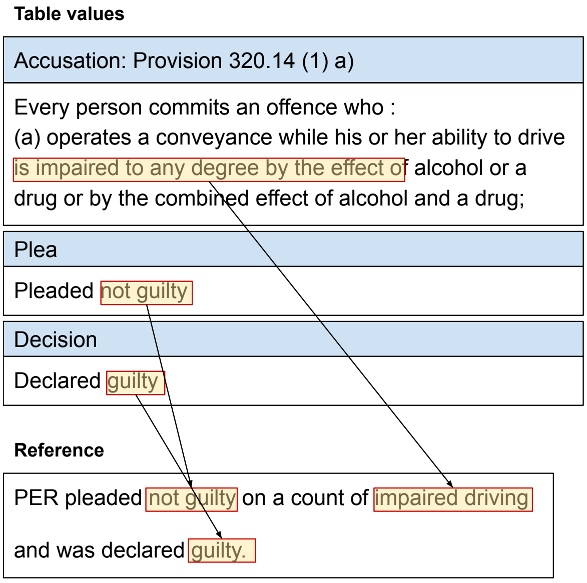


Table values

Accusation: Provision 320.14 (1) a)

Every person commits an offence who :
(a) operates a conveyance while his or her ability to drive is impaired to any degree by the effect of alcohol or a drug or by the combined effect of alcohol and a drug;

Plea

Pleaded not guilty

Decision

Declared guilty

Reference

PER pleaded not guilty on a count of impaired driving and was declared guilty.

Figure 5: An instance from *Plum2Text*' training set containing three table values with its associated reference. Text has been translated from French to English.

Input Data	accusation: 348 1) a) Introduction par effraction dans un dessein criminel: quiconque, selon le cas: s'introduit en un endroit par effraction avec l'intention d'y commettre un acte criminel; plaidoyer: coupable.
LSTM	l'accusé a plaidé coupable à trois chefs de trafic d'héroïne et un chef de possession en vue de trafic de cette drogue.
BARThez	PER a plaidé coupable à des accusations de s'être introduit par effraction dans une maison d'habitation , de s'être livré à des voies de fait, de s'être évadé, de s'être livré à des voies de fait, de s'être livré à des voies de fait et de s'être livré à des actes criminels.
***Criminel*BART**	PER plaide coupable à une accusation d'introduction par effraction dans une maison d'habitation avec l'intention d' y commettre un acte criminel.

Table 6: Example generations from the three models on the input data of provision 348 1) a) and a guilty plea. We can see that the LSTM is completely off-track (drug trafficking) while BARThez hallucinates several facts (assault and escaped from jail). *Criminel*BART contains the good provision (breaking and entering with the intent to commit a crime), but hallucinates "in a dwelling house".

C Generation Examples

Table 6 presents example generations from the three models given the input "provision 348 and a guilty plea".

D Unseen Provisions

- **445.1 (1) a)**: "*Causing unnecessary suffering. Every one commits an offence who wilfully causes or, being the owner, wilfully permits to be caused unnecessary pain, suffering or injury to an animal or a bird;*"
- **150**: "Illegally had in his possession for sale magazines that are obscene."
- **83.181**: "*Leaving Canada to participate in activity of terrorist group. Every person who leaves or attempts to leave Canada, or goes or attempts to go on board a conveyance with the intent to leave Canada, for the purpose of committing an act or omission outside Canada that, if committed in Canada, would be an offence under subsection 83.18(1) is guilty of an indictable offence and liable to imprisonment for a term of not more than 10 years.*"
- **241 (1) a)**: "*Counselling or aiding suicide. Everyone is guilty of an indictable offence and liable to imprisonment for a term of not more than 14 years who, whether suicide ensues or not, counsels a person to die by suicide or abets a person in dying by suicide;*"
- **811 a)**: "*Breach of recognizance. person bound by a recognizance under any of sections 83.3 and 810 to 810.2 who commits a breach of the recognizance is guilty of an indictable offence and is liable to imprisonment for a term of not more than four years;*"
- **467.12 (1)**: "*Commission of offence for criminal organization. Every person who commits an indictable offence under this or any other Act of Parliament for the benefit of, at the direction of, or in association with, a criminal organization is guilty of an indictable offence and liable to imprisonment for a term not exceeding fourteen years.*"
- **810.2**: "*Where fear of serious personal injury offence. Any person who fears on reasonable grounds that another person will commit a serious personal injury offence, as that expression is defined in section 752, may, with the consent of the Attorney General, lay an information before a provincial court judge, whether or not the person or persons in respect of whom it is feared that the offence will be committed are named.*"
- **172 (1) a)**: "*Corrupting children. Every person who, in the home of a child, participates in adultery or sexual immorality or indulges in habitual drunkenness or any other form of vice, and by doing so endangers the morals of the child or renders the home an unfit place for the child to be in, is guilty of an indictable offence and liable to imprisonment for a term of not more than two years;*"
- **320.14 (1) a)**: "*Operation while impaired. Everyone commits an offence who operates a conveyance while the person's ability to operate it is impaired to any degree by alcohol or a drug or by a combination of alcohol and a drug;*"

E Human Evaluation Datasheet

E.1 Paper and Supplementary Resources (Questions 1.1–1.3)

Question 1.1: Link to paper reporting the evaluation experiment. If the paper reports more than one experiment, state which experiment you're completing this sheet for. Or, if applicable, enter 'for preregistration.'

For preregistration.

Question 1.2: Link to website providing resources used in the evaluation experiment (e.g. system outputs, evaluation tools, etc.). If there isn't one, enter 'N/A'.

N/A.

Question 1.3: Name, affiliation and email address of person completing this sheet, and of contact author if different.

Will be completed upon acceptation.

E.2 System (Questions 2.1–2.5)

Question 2.1: What type of input do the evaluated system(s) take? Select all that apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

✓ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

□ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; **?**).

□ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

□ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

□ ***text: sentence***: a single sentence (or set of sentences).

□ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

□ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

□ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

□ ***text: other***: input is text but doesn't match any of the above *text:** categories.

□ ***speech***: a recording of speech.

□ ***visual***: an image or video.

□ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

□ ***control feature***: a feature or parameter specifically present to control a property of the output text, e.g. positive stance, formality, author style.

□ ***no input (human generation)***: human generation[8], therefore no system inputs.

□ ***other (please specify)***: if input is none of the above, choose this option and describe it.

Question 2.2: What type of output do the evaluated system(s) generate? Select all that apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

□ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

[8] We use the term 'human generation' where the items being evaluated have been created manually, rather than generated by an automatic system.

□ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; **?**).

□ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

□ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

✓ ***text: sentence***: a single sentence (or set of sentences).

□ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

□ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

□ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

□ ***text: other***: select if output is text but doesn't match any of the above *text:** categories.

□ ***speech***: a recording of speech.

□ ***visual***: an image or video.

□ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

□ ***human-generated 'outputs'***: manually created stand-ins exemplifying outputs.

□ ***other (please specify)***: if output is none of the above, choose this option and describe it.

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2? Occasionally, more than one of the options below may apply. If none match, select 'Other' and describe.

Check-box options (select all that apply):

□ ***content selection/determination***: selecting the specific content that will be expressed in the generated text from a representation of possible content. This could be attribute selection for REG (without the surface realisation step). Note that the output here is not text.

□ ***content ordering/structuring***: assigning an order and/or structure to content to be included in generated text. Note that the output here is not text.

□ ***aggregation***: converting inputs (typically *deep linguistic representations* or *shallow linguistic representations*) in some way in order to reduce redundancy (e.g. representations for 'they like swimming', 'they like running' → representation for 'they like swimming and running').

□ ***referring expression generation***: generating *text* to refer to a given referent, typically represented in the input as a set of attributes or a linguistic representation.

□ ***lexicalisation***: associating (parts of) an input representation with specific lexical items to be used in their realisation.

✓ ***deep generation***: one-step text generation from *raw/structured data* or *deep linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***surface realisation (SLR to text)***: one-step text generation from *shallow linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***feature-controlled text generation***: generation of text that varies along specific dimensions where the variation is controlled via *control features* specified as part of the input. Input is a non-textual representation (for feature-controlled text-to-text generation select the matching text-to-text task).

✓ ***data-to-text generation***: generation from *raw/structured data* which may or may not include some amount of content selection as part of the generation process. Output is likely to be *text:** or *multi-modal*.

□ ***dialogue turn generation***: generating a dialogue turn (can be a greeting or closing) from a representation of dialogue state and/or last turn(s), etc.

□ ***question generation***: generation of questions from given input text and/or knowledge base such that the question can be answered from the input.

- □ ***question answering***: input is a question plus optionally a set of reference texts and/or knowledge base, and the output is the answer to the question.
- ✓ ***paraphrasing/lossless simplification***: text-to-text generation where the aim is to preserve the meaning of the input while changing its wording. This can include the aim of changing the text on a given dimension, e.g. making it simpler, changing its stance or sentiment, etc., which may be controllable via input features. Note that this task type includes meaning-preserving text simplification (non-meaning preserving simplification comes under *compression/lossy simplification* below).
- □ ***compression/lossy simplification***: text-to-text generation that has the aim to generate a shorter, or shorter and simpler, version of the input text. This will normally affect meaning to some extent, but as a side effect, rather than the primary aim, as is the case in *summarisation*.
- □ ***machine translation***: translating text in a source language to text in a target language while maximally preserving the meaning.
- □ ***summarisation (text-to-text)***: output is an extractive or abstractive summary of the important/relevant/salient content of the input document(s).
- □ ***end-to-end text generation***: use this option if the single system task corresponds to more than one of tasks above, implemented either as separate modules pipelined together, or as one-step generation, other than *deep generation* and *surface realisation*.
- □ ***image/video description***: input includes *visual*, and the output describes it in some way.
- □ ***post-editing/correction***: system edits and/or corrects the input text (typically itself the textual output from another system) to yield an improved version of the text.
- □ ***other (please specify)***: if task is none of the above, choose this option and describe it.

Question 2.4: Input Language(s), or ‘N/A’.

French.

Question 2.5: Output Language(s), or ‘N/A’.

French.

E.3 Output Sample, Evaluators, Experimental Design

E.3.1 Sample of system outputs (or human-authored stand-ins) evaluated (Questions 3.1.1–3.1.3)

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment? Answer should be an integer.

89.

Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment? If none match, select ‘Other’ and describe.

Multiple-choice options (select one):

- ○ ***by an automatic random process from a larger set***: outputs were selected for inclusion in the experiment by a script using a pseudo-random number generator; don’t use this option if the script selects every nth output (which is not random).
- ○ ***by an automatic random process but using stratified sampling over given properties***: use this option if selection was by a random script as above, but with added constraints ensuring that the sample is representative of the set of outputs it was selected from, in terms of given properties, such as sentence length, positive/negative stance, etc.
- ○ ***by manual, arbitrary selection***: output sample was selected by hand, or automatically from a manually compiled list, without a specific selection criterion.
- ✓ ***by manual selection aimed at achieving balance or variety relative to given properties***: selection by hand as above, but with specific selection criteria, e.g. same number of outputs from each time period.

○ ***Other (please specify)***: if selection method is none of the above, choose this option and describe it.

Question 3.1.3: What is the statistical power of the sample size?

Following the methodology of Card et al. (2020), we obtained a statistical power of 1.0 on the output sample w.r.t the automatic evaluation metrics, the two best performing models (BARThez and *Criminel*BART). We used their online script to estimate the statistical power.

E.3.2 Evaluators (Questions 3.2.1–3.2.4)

Question 3.2.1: How many evaluators are there in this experiment? Answer should be an integer.

Three.

Question 3.2.2: What kind of evaluators are in this experiment? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***experts***: participants are considered domain experts, e.g. meteorologists evaluating a weather forecast generator, or nurses evaluating an ICU report generator.

□ ***non-experts***: participants are not domain experts.

✓ ***paid (including non-monetary compensation such as course credits)***: participants were given some form of compensation for their participation, including vouchers, course credits, and reimbursement for travel unless based on receipts.

□ ***not paid***: participants were not given compensation of any kind.

□ ***previously known to authors***: (one of the) researchers running the experiment knew some or all of the participants before recruiting them for the experiment.

✓ ***not previously known to authors***: none of the researchers running the experiment knew any of the participants before recruiting them for the experiment.

✓ ***evaluators include one or more of the authors***: one or more researchers running the experiment was among the participants.

□ ***evaluators do not include any of the authors***: none of the researchers running the experiment were among the participants.

□ ***Other*** (fewer than 4 of the above apply): we believe you should be able to tick 4 options of the above. If that's not the case, use this box to explain.

Question 3.2.3: How are evaluators recruited?

Evaluators (excluding one or more of the authors) were recruited by word of mouth, and have been interviewed prior to conduct the experiment.

Question 3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself?

First, the evaluators have been introduced to the task of data-to-text generation. They then have been introduced to the dataset under study. They learned from an annotation guideline and have practiced on 5 examples before conducting the whole experiment. Evaluators did not need legal training since they had background knowledge on the domain.

Question 3.2.5: What other characteristics do the evaluators have, known either because these were qualifying criteria, or from information gathered as part of the evaluation?

Evaluators have been selected based on their educational level (2 years in law school) and their interest in criminal law.

E.3.3 Experimental design (Questions 3.3.1–3.3.8)

Question 3.3.1: Has the experimental design been preregistered? If yes, on which registry?

No.

Question 3.3.2: How are responses collected? E.g. paper forms, online survey tool, etc.

The answers were collected using a customized version of Prodigy[9], hosted on Amazon Web Services.

Question 3.3.3: What quality assurance methods are used? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***evaluators are required to be native speakers of the language they evaluate***: mechanisms are in place to ensure all participants are native speakers of the language they evaluate.

□ ***automatic quality checking methods are used during/post evaluation***: evaluations are checked for quality by automatic scripts during or after evaluations, e.g. evaluators are given known bad/good outputs to check they're given bad/good scores on MTurk.

✓ ***manual quality checking methods are used during/post evaluation***: evaluations are checked for quality by a manual process during or after evaluations, e.g. scores assigned by evaluators are monitored by researchers conducting the experiment.

□ ***evaluators are excluded if they fail quality checks (often or badly enough)***: there are conditions under which evaluations produced by participants are not included in the final results due to quality issues.

□ ***some evaluations are excluded because of failed quality checks***: there are conditions under which some (but not all) of the evaluations produced by some participants are not included in the final results due to quality issues.

□ ***none of the above***: tick this box if none of the above apply.

□ ***Other (please specify)***: use this box to describe any other quality assurance methods used during or after evaluations, and to provide additional details for any of the options selected above.

Question 3.3.4: What do evaluators see when carrying out evaluations? Link to screenshot(s) and/or describe the evaluation interface(s).

When carrying out evaluations, evaluators see the input data as well as three generations from three different models. They do not know which generation corresponds to which model. They then provide a score for each generation independently.

3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

□ ***evaluators have to complete each individual assessment within a set time***: evaluators are timed while carrying out each assessment and cannot complete the assessment once time has run out.

□ ***evaluators have to complete the whole evaluation in one sitting***: partial progress cannot be saved and the evaluation returned to on a later occasion.

✓ ***neither of the above***: Choose this option if neither of the above are the case in the experiment.

□ ***Other (please specify)***: Use this space to describe any other way in which time taken or number of sessions used by evaluators is controlled in the experiment, and to provide additional details for any of the options selected above.

[9] https://prodi.gy/

3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *before* starting on their assessments, either during or after training.

□ ***evaluators are told they can ask any questions during the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *during* their assessments.

□ ***evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box***: evaluators are explicitly asked to provide feedback and/or comments about the experiment *after* their assessments, either verbally or in written form.

□ ***None of the above***: Choose this option if none of the above are the case in the experiment.

□ ***Other (please specify)***: use this space to describe any other ways you provide for evaluators to ask questions or provide feedback.

3.3.7: What are the experimental conditions in which evaluators carry out the evaluations? If none match, select 'Other' and describe.

Multiple-choice options (select one):

✓ ***evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.***: evaluators are given access to the tool or form specified in Question 3.3.2, and subsequently choose where to carry out their evaluations.

○ ***evaluation carried out in a lab, and conditions are the same for each evaluator***: evaluations are carried out in a lab, and conditions in which evaluations are carried out *are* controlled to be the same, i.e. the different evaluators all carry out the evaluations in identical conditions of quietness, same type of computer, same room, etc. Note we're not after very fine-grained differences here, such as time of day or temperature, but the line is difficult to draw, so some judgment is involved here.

○ ***evaluation carried out in a lab, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a lab, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***evaluation carried out in a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out in a real-life situation, i.e. one that would occur whether or not the evaluation was carried out (e.g. evaluating a dialogue system deployed in a live chat function on a website), and conditions in which evaluations are carried out *are* controlled to be the same.

○ ***evaluation carried out in a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation (but not actually a real-life situation), e.g. user-testing a navigation system where the destination is part of the evaluation design, rather than chosen by the user. Conditions in which evaluations are carried out *are* controlled to be the same.

○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***Other (please specify)***: Use this space to provide additional, or alternative, information about the conditions in which evaluators carry out assessments, not covered by the options above.

3.3.8: Unless the evaluation is carried out at a place of the evaluators' own choosing, briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

N/A.

E.4 Quality Criterion *n* – Definition and Operationalisation

E.4.1 Quality criterion properties (Questions 4.1.1–4.1.3)

Question 4.1.1: What type of quality is assessed by the quality criterion?

Multiple-choice options (select one):

✓ ***Correctness***: select this option if it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality). E.g. for Grammaticality, outputs are (maximally) correct if they contain no grammatical errors; for Semantic Completeness, outputs are correct if they express all the content in the input.

○ ***Goodness***: select this option if, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for two outputs which is better and which is worse. E.g. for Fluency, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.

○ ***Features***: choose this option if, in terms of property X captured by the criterion, outputs are not generally better if they are more X, but instead, depending on evaluation context, more X may be better or less X may be better. E.g. outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

Question 4.1.2: Which aspect of system outputs is assessed by the quality criterion?

Multiple-choice options (select one):

○ ***Form of output***: choose this option if the criterion assesses the form of outputs alone, e.g. Grammaticality is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.

✓ ***Content of output***: choose this option if the criterion assesses the content/meaning of the output alone, e.g. Meaning Preservation only assesses output content; two sentences can be considered to have the same meaning, but differ in form.

○ ***Both form and content of output***: choose this option if the criterion assesses outputs as a whole, not just form or just content. E.g. Coherence is a property of outputs as a whole, either form or meaning can detract from it.

Question 4.1.3: Is each output assessed for quality in its own right, or with reference to a system-internal or external frame of reference?

Multiple-choice options (select one):

○ ***Quality of output in its own right***: choose this option if output quality is assessed without referring to anything other than the output itself, i.e. no system-internal or external frame of reference. E.g. Poeticness is assessed by considering (just) the output and how poetic it is.

✓ ***Quality of output relative to the input***: choose this option if output quality is assessed relative to the input. E.g. Answerability is the degree to which the output question can be answered from information in the input.

○ ***Quality of output relative to a system-external frame of reference***: choose this option if output quality is assessed with reference to system-external information, such as a knowledge base, a person's individual writing style, or the performance of an embedding system. E.g. Factual Accuracy assesses outputs relative to a source of real-world knowledge.

E.4.2 Evaluation mode properties (Questions 4.2.1–4.2.3)

Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criteria, i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

Question 4.2.1: Does an individual assessment involve an objective or a subjective judgment?

Multiple-choice options (select one):

✓ ***Objective***: Examples of objective assessment include any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. Repeated assessments of the same output with an objective-mode evaluation method always yield the same score/result.

○ ***Subjective***: Subjective assessments involve ratings, opinions and preferences by evaluators. Some criteria lend themselves more readily to subjective assessments, e.g. Friendliness of a conversational agent, but an objective measure e.g. based on lexical markers is also conceivable.

Question 4.2.2: Are outputs assessed in absolute or relative terms?

Multiple-choice options (select one):

○ ***Absolute***: choose this option if evaluators are shown outputs from a single system during each individual assessment.

✓ ***Relative***: choose this option if evaluators are shown outputs from multiple systems at the same time during assessments, typically ranking or preference-judging them.

Question 4.2.3: Is the evaluation intrinsic or extrinsic?

Multiple-choice options (select one):

○ ***Intrinsic***: Choose this option if quality of outputs is assessed *without* considering their *effect* on something external to the system, e.g. the performance of an embedding system or of a user at a task.

✓ ***Extrinsic***: Choose this option if quality of outputs is assessed in terms of their *effect* on something external to the system such as the performance of an embedding system or of a user at a task.

E.4.3 Response elicitation (Questions 4.3.1–4.3.11)

Question 4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if criterion not named.

Legal accuracy.

Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.

We define legal accuracy as being a text that respectfully captures the input data w.r.t the criminal code, the plea and the verdict. In most cases, legal accuracy w.r.t the criminal code is the hardest part of the task for neural networks.

Question 4.3.3: Size of scale or other rating instrument (i.e. how many different possible values there are). Answer should be an integer or 'continuous' (if it's not possible to state how many possible responses there are). Enter 'N/A' if there is no rating instrument.

10.

Question 4.3.4: List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.

1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Question 4.3.5: How is the scale or other rating instrument presented to evaluators? If none match, select 'Other' and describe.

Multiple-choice options (select one):

○ ***Multiple-choice options***: choose this option if evaluators select exactly one of multiple options.

○ ***Check-boxes***: choose this option if evaluators select any number of options from multiple given options.

○ ***Slider***: choose this option if evaluators move a pointer on a slider scale to the position corresponding to their assessment.

○ ***N/A (there is no rating instrument)***: choose this option if there is no rating instrument.

✓ ***Other (please specify)***: choose this option if there is a rating instrument, but none of the above adequately describe the way you present it to evaluators. Use the text box to describe the rating instrument and link to a screenshot.

Due to the limitations of Prodigy regarding their slider component (only one per page), we used a free-form text box. Since we have few, highly skilled evaluators, it was not a problem collecting data.

Question 4.3.6: If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.

N/A.

Question 4.3.7: What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?

Do subsequent generations capture the data from the docket file? Rate on a scale of 1 to 10.

Question 4.3.8: Form of response elicitation. If none match, select 'Other' and describe.

Multiple-choice options (select one):[10]

○ ***(dis)agreement with quality statement***: Participants specify the degree to which they agree with a given quality statement by indicating their agreement on a rating instrument. The rating instrument is labelled with degrees of agreement and can additionally have numerical labels. E.g. *This text is fluent — 1=strongly disagree...5=strongly agree*.

○ ***direct quality estimation***: Participants are asked to provide a rating using a rating instrument, which typically (but not always) mentions the quality criterion explicitly. E.g. *How fluent is this text? — 1=not at all fluent...5=very fluent*.

○ ***relative quality estimation (including ranking)***: Participants evaluate two or more items in terms of which is better. E.g. *Rank these texts in terms of fluency*; *Which of these texts is more fluent?*; *Which of these items do you prefer?*.

✓ ***counting occurrences in text***: Evaluators are asked to count how many times some type of phenomenon occurs, e.g. the number of facts contained in the output that are inconsistent with the input.

○ ***qualitative feedback (e.g. via comments entered in a text box)***: Typically, these are responses to open-ended questions in a survey or interview.

○ ***evaluation through post-editing/annotation***: Choose this option if the evaluators' task consists of editing or inserting annotations in text. E.g. evaluators may perform error correction and edits are then automatically measured to yield a numerical score.

○ ***output classification or labelling***: Choose this option if evaluators assign outputs to categories. E.g. *What is the overall sentiment of this piece of text? — Positive/neutral/negative.*

○ ***user-text interaction measurements***: choose this option if participants in the evaluation experiment interact with a text in some way, and

[10] Explanations adapted from Howcroft et al. (2020).

measurements are taken of their interaction. E.g. reading speed, eye movement tracking, comprehension questions, etc. Excludes situations where participants are given a task to solve and their performance is measured which comes under the next option.

○ ***task performance measurements***: choose this option if participants in the evaluation experiment are given a task to perform, and measurements are taken of their performance at the task. E.g. task is finding information, and task performance measurement is task completion speed and success rate.

○ ***user-system interaction measurements***: choose this option if participants in the evaluation experiment interact with a system in some way, while measurements are taken of their interaction. E.g. duration of interaction, hyperlinks followed, number of likes, or completed sales.

○ ***Other (please specify)***: Use the text box to describe the form of response elicitation used in assessing the quality criterion if it doesn't fall in any of the above categories.

Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? State if no scores reported.

Macro-averages are computed from numerical scores to provide summary, per-system results.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.

What to enter in the text box: A list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state 'None'. None.

Question 4.3.11: Has the inter-annotator and intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used, and what are the agreement scores?

Krippendorff's alpha is used to measure inter-annotator agreement. Krippendorff's alpha is of 0.84.

F Ethics

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

No.

Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions/)? If yes, describe data and state how addressed.

No.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/)? If yes, describe data and state how addressed.

No.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.

No.