Modeling Subjective Assessments of Guilt in Newspaper Crime Narratives

Elisa Kreiss[1,*]   **Zijian Wang**[2,*]   **Christopher Potts**[1,2]

[1]Department of Linguistics   [2]Symbolic Systems Program
Stanford University
{ekreiss, zijwang, cgpotts}@stanford.edu

Abstract

Crime reporting is a prevalent form of journalism with the power to shape public perceptions and social policies. How does the language of these reports act on readers? We seek to address this question with the **SuspectGuilt Corpus** of annotated crime stories from English-language newspapers in the U.S. For Suspect-Guilt, annotators read short crime articles and provided text-level ratings concerning the guilt of the main suspect as well as span-level annotations indicating which parts of the story they felt most influenced their ratings. SuspectGuilt thus provides a rich picture of how linguistic choices affect subjective guilt judgments. We use SuspectGuilt to train and assess predictive models which validate the usefulness of the corpus, and show that these models benefit from genre pretraining and joint supervision from the text-level ratings and span-level annotations. Such models might be used as tools for understanding the societal effects of crime reporting.

1 Introduction

News outlets around the world routinely report on crimes and alleged crimes, ranging from petty misdemeanors to large-scale international criminal conspiracies. Each of these reports will frame events in ways that shape reader perceptions, and these perceptions will in turn shape public perception of how much crime there is, who is responsible for crime, and what policy decisions should be made to address crime. It is therefore important to understand how the language in these reports acts on readers, and there is clear value in developing NLP models that approximate these reader perceptions at a large scale, as a tool for estimating the aggregate effects of crime reporting on society.

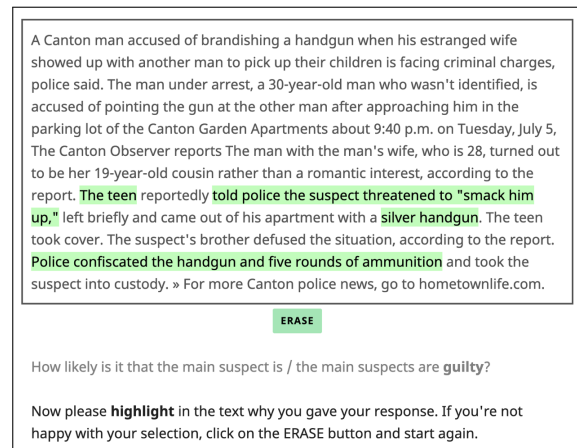To begin to address these needs, we present the **SuspectGuilt Corpus** of annotated crime stories

Figure 1: The SuspectGuilt corpus highlighting interface. After participants responded to a question about the guilt of the main suspect in the report, they completed this highlighting phase intended to provide insights into how they took themselves to be reasoning about the text. SuspectGuilt contains 1.8K stories with at least 5 participants responding to each.

from English-language newspapers in the U.S.[1] Each story in the corpus is multiply-annotated with participants' assessments (on a continuous scale) of the guilt of the main suspect(s) and of the author's belief in the guilt of the suspect(s). In addition, for each of these guilt-rating questions, the participants highlighted the spans of text in the story that they felt contributed to their decision (Figure 1). These additional annotations provide a window into the language that participants took themselves to be attending to as part of their personal verdicts, and thus they are especially useful for understanding how authors' low-level linguistic choices feed into readers' overall judgments.

We also explore a range of methods for developing predictive models on the basis of SuspectGuilt annotations which exemplify the usefulness of the

---

[1]https://github.com/zijwang/modeling_guilt

resource. Our models are built on top of pretrained BERT parameters. In the simplest case, we learn to predict the author or subject guilt ratings without any other supervision. This basic model is improved if it is jointly trained on the guilt ratings and the span-level annotations that SuspectGuilt provides, which helps to quantify the value of these low-level linguistic annotations. In addition, we explore unsupervised pretraining on a modestly-sized unlabeled corpus of crime stories, finding that it too increases the effectiveness of SuspectGuilt models.

The span-level annotations offer new opportunities for analysis as well. Using the Integrated Gradients method of Sundararajan et al. (2017), we identify the token-level features that our models rely on when trained without span-level supervision, and we compare this to the span-level annotations provided by SuspectGuilt. Overall, the correspondence between the two is not high, which explains why the span-level objective helps our models and suggests that the document-level ratings alone might not suffice to yield models that attend to texts in the same ways that humans do.

2 Related Work

Our work draws on prior research into the relationship between language and assessments of guilt, as well as work seeking to jointly model text-level and token-level annotations using neural networks.

2.1 Predicting Guilt

The challenge of predicting guilt judgments from text sources has not yet received much attention. However, Fausey and Boroditsky (2010) show that using agentive language increases blame and financial liability judgments people make. Their results suggest that even subtle linguistic changes in crime reports will shape people's judgments of the events. More recent work has focused on predicting guilt verdicts from the Supreme Courts in the Philippines (Virtucio et al., 2018) and Thailand (Kowsrihawat et al., 2018) on the basis of presented facts and legal texts. Kowsrihawat et al. employ a recurrent neural network with attention to make these predictions. These findings are for courtroom verdicts based on legal texts, and thus they are a useful complement to SuspectGuilt, which provides subjective guilt judgments based on crime reporting.

2.2 Veridicality Markers

We use the label 'veridicality markers' to informally identify a large class of lexical items that includes hedges, evidentials, and other markers of (un)certainty. Analysis of the span-level annotations in SuspectGuilt shows that veridicality markers play an out-sized role in shaping people's judgments of guilt. The annotations are dominated not only by conventionalized devices like *allegedly*, *suspect*, and *according to*, but also by more context-specific locutions like *police say* and *arrest*.

There is extensive prior literature on how veridicality markers affect the perceptions of the speaker and proposition (Erickson et al., 1978; Durik et al., 2008; Bonnefon and Villejoubert, 2006; Rubin, 2007; Jensen, 2008; Ferson et al., 2015). These studies suggest such markers affect people's judgments of credibility in differing ways. For example, an increase in the number of hedges decreases the credibility of witness reports (Erickson et al., 1978) but increases the trustworthiness of journalists and scientists (Jensen, 2008). Additionally, the interpretation of hedges is context dependent (Bonnefon and Villejoubert, 2006; Durik et al., 2008; Ferson et al., 2015) and show high individual variation (Rubin, 2007; Ferson et al., 2015).

Similarly, attitude predications like *X reported S* can be used to reduce commitment, but they can also be used to provide evidence in favor of *S* (Simons 2007; de Marneffe et al. 2012; White and Rawlins 2018; White et al. 2018). Stone (1994) and von Fintel and Gillies (2010) address similar uses of epistemic modal verbs. These findings show how complex these markers are pragmatically and highlight the value of usage-based studies of them.

2.3 Span-Level Supervision

BERT models (Devlin et al., 2019) define an output representation for every token-level input (see also Vaswani et al. 2017). The parameters of these models can be fine-tuned in many ways (Lee et al., 2020; Mosbach et al., 2020). Our models combine text-level prediction with sequence modeling; the supervision signals come from the guilt judgments and span highlighting in the SuspectGuilt corpus. This basic model structure has been used in a wide variety of settings before. What is perhaps special about our use of it is that the two levels of annotation each provide evidence about the other; the highlighting can be seen as guiding the regression model to pay attention to certain words, and the

regression label is likely to create helpful biases for particular token-level classifications. Rei and Søgaard (2019) define models that similarly make use complementary tasks. This is also conceptually very similar to the token-level supervision in the debiasing model of Pryzant et al. (2020). However, while their token-level labels come from a fixed lexicon, ours were created in their linguistic context with a particular set of guilt-related issues in mind.

3 Data

The SuspectGuilt corpus is a resource to investigate how the language of crime reports affects readers. This section describes the data collection and annotation process. We provide qualitative and quantitative analyses of SuspectGuilt that exemplify its usefulness for psycholinguistic investigations and NLP applications.

3.1 Data Collection

The SuspectGuilt corpus is derived from a dataset of crime-related newspaper stories from regional, English-language newspapers in the U.S. We chose to focus on such stories because they are generally brief and self-contained. By contrast, crime-related stories from major news outlets tend to involve public figures, political issues, and important global events, and readers' prior exposure to the issues might affect their judgments in unpredictable ways.

Inspired by Davani et al. (2019), we collected our corpus from Patch.com. The Patch dataset contains independent, hyper-local news articles compiled from local news sites. We crawled all stories in the "Crime & Safety" section for all news up through December 2019, yielding 474k news stories from 1,226 communities in the U.S. We then filtered this collection to just stories with (1) at most 300 words and (2) at least 4 of the following word-stems: `suspect*`, `alleg*`, `arrest*`, `crim*`, `accus*`[2]. In addition, we filtered out stories that either have the same title, for which we only keep one copy, or are collections of multiple reports, e.g., records of incidents. As a post-processing step, we removed phone numbers and Patch.com advertisements. The final collection has 4.2K stories, of which we selected 1,957 for annotation.

---

[2]The word-stems were chosen to maximize the retrieval of news stories that report on criminal acts where a suspect has been identified but that still communicate uncertainty about the case.
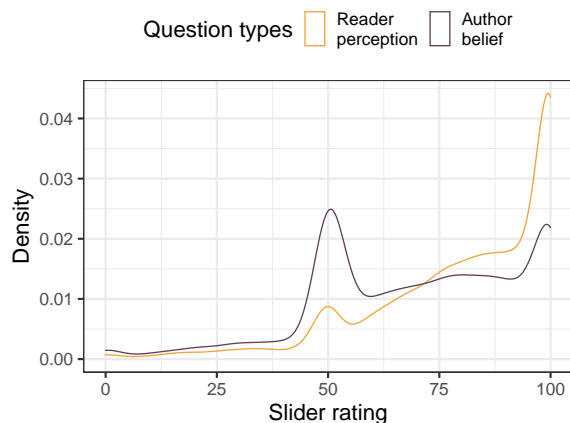

Figure 2: Slider rating density distribution for the *Reader perception* and *Author belief* questions.

3.2 Annotation Effort

For the annotation phase of SuspectGuilt, participants were recruited on Amazon's Mechanical Turk and asked to read five stories and respond to three questions about them:

1. *Reader perception*: "How likely is it that the main suspect is / the main suspects are guilty?"

2. *Author belief*: "How much does the author believe that the main suspect is / the main suspects are guilty?"

3. An attention check question, such as "How likely is it that this story contains more than five words?"

Responses were collected on a continuous slider, coded as ranging from 0 (very unlikely) to 1 (very likely). After submitting the slider response for each question, participants were asked to "highlight in the text why [they] gave [their] response". They additionally had the option to opt out of the slider response by indicating that the question didn't apply to the story. Stories with more than 30% of "Doesn't apply" responses were excluded from the corpus, yielding 1,821 unique news reports.

Guilt judgments are subjective and known to be highly variable (Section 2.2), and we expect the span-level highlighting to be even more variable. To accommodate this natural variation, we had multiple participants rate each story. Every story was annotated at least 5 times, and after excluding "Doesn't apply" responses, 99.2% of the stories still have 5 annotations or more for the *Reader perception* question and 86.7% for the *Author belief*
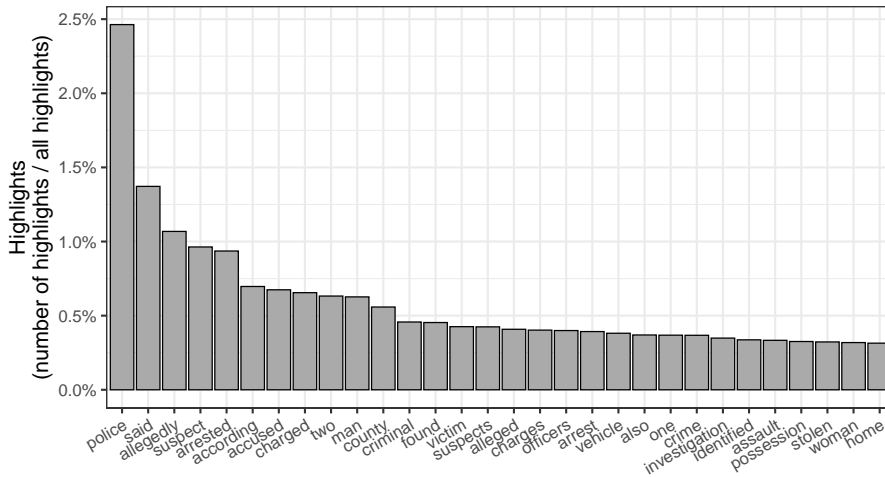
58

Figure 3: The 30 most highlighted words across questions.

question. For our analyses and modeling in this paper, we generally average these annotations, but the corpus supports work at finer-grained levels. Our appendices include additional details, including screenshots of the annotation interface, exclusion criteria for participants, and aggregated participant demographics.

3.3 Text-Level Annotations

Figure 2 shows the distribution of responses for the *Reader perception* and *Author belief* questions. Both distributions are skewed towards the middle and maximum portions of the slider scale. Relatively few participants chose ratings in the "very unlikely" range, which potentially reflects underlying biases about news reporting: readers expect suspects mentioned in these stories to be guilty. We also begin to see differences between the two questions. While *Reader perception* ratings are rather skewed to the maximum portion of the scale, *Author belief* responses are concentrated around the center. This already suggests a disconnect between what readers believe about the suspect's guilt more generally and what readers believe about the author's beliefs. The cluster around the center also suggests that participants feel uncertainty, especially in the *Author belief* case. The clustering might also reflect a presumption that journalists will seek to appear unbiased.

We find high levels of interannotator agreement for both the *Reader perception* and *Author belief* questions. The mean squared error (MSE) for each story is lower for the *Reader perception* question (mean MSE = 0.0313) than *Author belief* (mean

MSE = 0.0410). To provide some context for these numbers, we also calculated them after first shuffling all ratings. The MSE for this setting is 0.0443 for *Author belief* and 0.0353 for *Reader perception*. Both are significantly higher than their non-shuffled counterparts according to a Welch Two Sample t-test ($p < 0.0001$).

3.4 Span-Level Annotations

When highlighting text spans, participants primarily marked passages shorter than 200 characters (approximately 33 words). *Author belief* highlights tended to be shorter than those for *Reader perception*. Overall, highlights had a length between 1 and 1,717 characters (about 286 words). (A highlight here is defined as a consecutive mark without a non-highlighted character in between. If a participant highlighted two passages that are directly connected, they count as one highlighting.)

We would like to estimate agreement levels for span highlighting as well. Because our stories have varying numbers of annotations, we cannot calculate a Fleiss kappa value for this problem. Krippendorff's alpha is a standard test that can accommodate this kind of variation, but its symmetric treatment of highlighting and non-highlighting is problematic since only 15% of the tokens are highlighted.[3] Nonetheless, to provide some insight into how alike our participants were in their highlighting behavior, we compared the percentage of anno-

---

[3]Due to this inequality, the random baseline in Krippendorff's alpha (which is computed by shuffling each story's highlights) is disproportionately strong. The highlighting data still achieves a positive Krippendorff's alpha of 0.16 for *Reader perception* and 0.08 for *Author belief*.

Figure 4: Proportion of token selections by frequency. Words that received the most highlights overall (see Figure 3) are presented in red, other words in grey. By chance, words would be highlighted 14.88% of the time, indicated by the dashed grey line. Words that are highlighted more often than predicted by chance are above this line, suggesting that they take on an important role in annotators' judgments.

tators who highlighted each character with a random baseline. The random baseline highlights were created by randomly shuffling the underlying highlight distribution for each annotation. We find that it was more likely that at least half of the annotators considered a token as important in the actual data as opposed to the random baseline (Welch Two Sample t-test: $p < 0.0001$).

Token-level analysis of the highlighted spans reveals many connections with the markers of veridicality discussed in Section 2.2. Figure 3 shows the most highlighted words across the two guilt questions.[4] The list is dominated by conventionalized devices for signaling lack of commitment in newspaper reporting (e.g., forms of *allege*), devices for shifting attribution to others (e.g., *said*, *accused*), and genre-specific words that play into how we assess evidence in criminal contexts (e.g., *accused*, *charged*, *investigation*).

However, as we might expect, the number of times a word is highlighted highly correlates with its frequency ($r = 0.97$). Figure 4 brings out this relationship. The x-axis is token frequency, and the y-axis gives the proportion of tokens for a word that were highlighted. (For example, if a word appeared 100 times and was highlighted 10 of those times, it would appear at 0.1 on the y-axis.) We excluded words with a frequency below 25, since these tend to get exaggerated proportions. The words from Figure 3 are displayed in red and are

highly frequent, and they are also the words with the highest highlighting proportion for their frequency, suggesting that these patterns are robust. Many of the other proportionally frequently highlighted words fall into the same categories as those in Figure 3: forms of *confess*, *eyewitnesses*, words picking out devices that provide evidence, and so forth. Words which were highlighted less than expected by chance (i.e., below the dashed grey line) rather reference meta-information of the news stories, such as *google*, *newsletter*, *shutterstock*, and *map*. In sum, the highlighting patterns seem aligned with the linguistic picture outlined in Section 2.2.

Figure 5 seeks to add a further dimension to this analysis. Thus far, we have ignored the distinction between the two guilt-rating questions, *Reader perception* and *Author belief*. The two questions are semantically quite different and might even come apart in some cases. For example, a reader might attend only to the evidence presented in a text and arrive at a high guilt-rating of their own, while ignoring clear indicators that the author wishes to remain non-committal about the origin or strength of that evidence. Kreiss et al. (2019) found that hedges affect responses of *Author belief* but not *Reader perception*, suggesting that the use of words like *allegedly* affects reader's perception about the author's beliefs but not their general guilt perception. This seems to be reflected in the selection data as well. In Figure 5, we give the words with the largest differences between the two guilt questions.

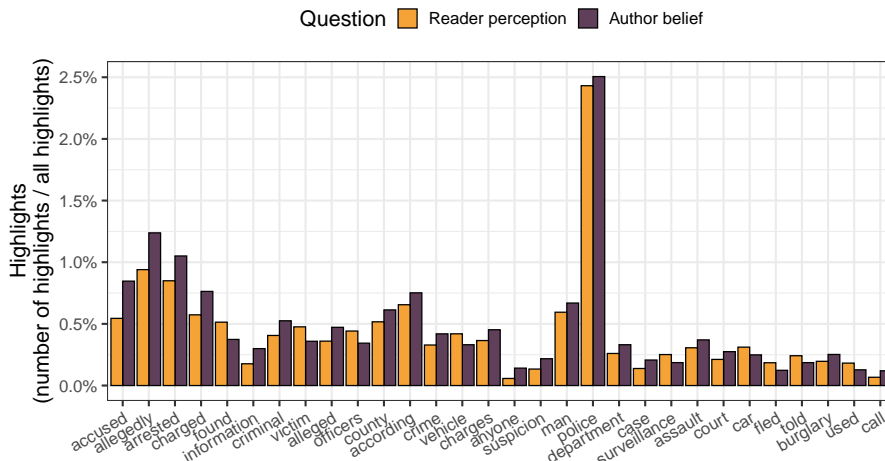[4]Punctuation and stopwords taken from the *tm: Text mining package* in R were excluded for this analysis.

Figure 5: Words with the largest highlighting difference between the two guilt questions.

Conventionalized devices like these hedges, which signal lack of commitment in reporting, become even more prominent in the *Author belief* condition. This supports Kreiss et al.'s earlier findings of the relevance of these words for *Author belief* and not *Reader perception*, and further suggests that readers appear to have some metalinguistic awareness for this difference.

4 Models

This section summarizes the family of models we consider in this work. All of them begin with BERT. We explore models with and without additional unsupervised pretraining on crime stories. We build regression models on top of these parameters using just the CLS token, which is the initial token in all BERT input sequences and is often taken to provide an aggregate sequence representation, as well as mean-pooling over all the final output states, and we additionally define extensions for predicting token-level highlighting.

4.1 Guilt Ratings

BERT (Devlin et al., 2019) is a Transformer-based architecture (Vaswani et al., 2017) that is usually trained jointly to do masked language modeling and next sentence prediction. The inputs are sequences of tokens $[x_0, \ldots, x_n]$, with x_0 designated as CLS and x_n designated as SEP. BERT maps these inputs to a sequence of output representations $[h_0, \ldots, h_n]$.

Our two rating categories, *Reader perception* and *Author belief*, define two separate tasks. We model them separately. For each, the core regres-

sion model is given by $hW_r + b_r$, where W_r is a vector of weights, b_r is a bias term, and h is derived from the states $[h_0, \ldots, h_n]$. In the CLS-based approach, $h = h_0$. In the mean-pooling approach, $h = \mathbf{mean}([h_0, \ldots, h_n])$.

The individual regression models are trained using a mean squared error (MSE) loss:

$$J_r(\theta_r) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \|H_{\theta_r}(x_i) - y_i^r\|^2 \quad (1)$$

Here, m is the number of examples, θ_r represents all the parameters of BERT plus our new task-specific parameters W_r and b_r, y_i^r is the true label for example x_i, and $H_{\theta_r}(x_i)$ is the prediction of the model for example x_i.

4.2 Genre Pretraining

BERT was trained on the BookCorpus (Zhu et al., 2015) and Wikipedia. It often performs well on tasks involving very different data, but any domain shift has the potential to lower performance, and crime stories are a specialized genre. Previous work has shown that in-domain continued pretraining is often beneficial for end-task performance in such situations (e.g. Han and Eisenstein, 2019; Gururangan et al., 2020). We thus evaluate models with and without pretraining on unlabeled crime stories. For this, we use the unlabeled portion of the dataset described in Section 3.1.

4.3 Span Highlighting

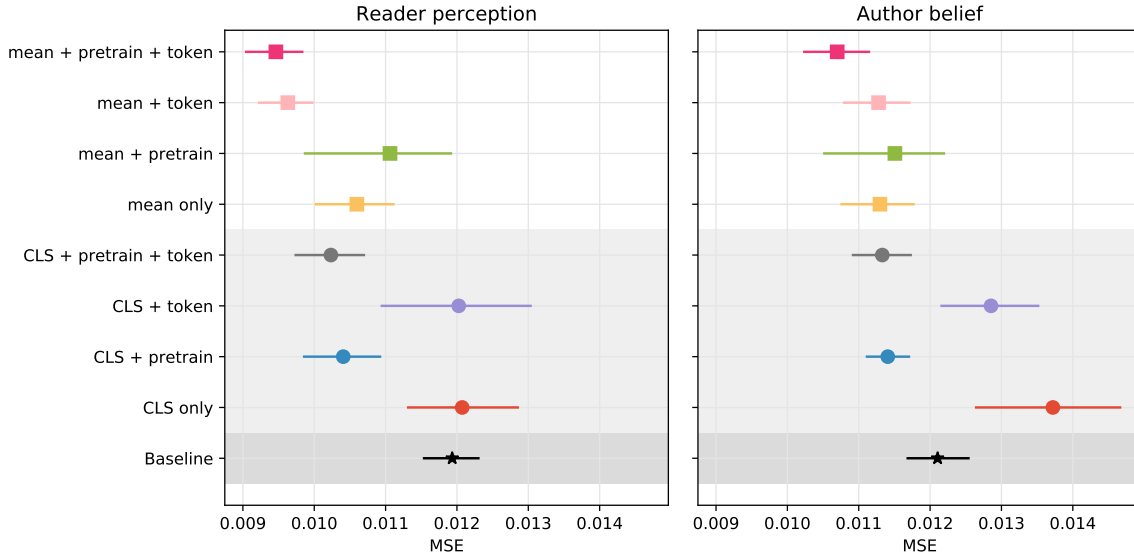We want to understand how authors' low-level linguistic choices affect readers' judgments of suspect

Figure 6: MSE (lower is better) for predicting guilt ratings for the *Reader perception* and *Author belief* questions, with bootstrapped 95% confidence intervals from 20 runs per model. 'CLS' models use just the `CLS` token for the regression, whereas 'mean' models average all the output heads (Section 4.1). 'pretrain' refers to genre pretraining (Section 4.2), and 'token' refers to token-level supervision from highlighting (Section 4.3).

guilt. To do this, we utilize the span-level annotation in SuspectGuilt. Annotations are coded as 1 if the token was highlighted, and 0 otherwise. We merge the annotations of each news story to form a supplemental regression task, where the target value is the mean of the annotations. We use the output representation of each token from BERT and apply a linear regression similar to (1):

$$J_t(\theta_t) = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\frac{1}{2}\|H_{\theta_t}(x_{ij}) - y_{ij}^t\|^2 \qquad (2)$$

Here, m is the number of examples, n is the number of tokens, and x_{ij} and y_{ij} stand for the jth token in example i, with corresponding token label y_{ij}^t. θ_t denotes all the BERT parameters plus token-level regression parameters W_t and b_t, and $H_{\theta_t}(x_{ij})$ is the prediction of the model for x_{ij}.

Our problem formulation might be taken to more naturally suggest a logistic regression. However, we opted for a linear regression objective instead, in the hopes that this would better capture not just the probability that a token is important, but also *how important these tokens are*. The linear regression performed better in our evaluations, though the improvements over the logistic were modest.

	dev	test
BERT-based	2.224	2.223
Genre Pretrained	0.884	0.887

Table 1: Losses with and without genre pretraining.

4.4 Joint Objective

The joint loss is a combination of the guilt-rating and span-highlighting objectives (1) and (2):

$$J = J_r(\theta_r) + \lambda J_t(\theta_t) \qquad (3)$$

where λ is a ratio of the losses that can be tuned.

5 Experiments

In this section, we report the evaluation procedure for the models described in Section 4. The results underline the usefulness of genre-pretraining and the rich annotations in the SuspectGuilt corpus.

5.1 Methods

We use the BERT-base uncased parameters for all of our experiments.

As discussed in Section 4.2, we performed pretraining with the $\approx$470K unlabeled articles from the dataset described in Section 3.1. We split the dataset into 80% training, 10% dev, and 10% test sets. Additional details are in Appendix B.1. Table 1 summarizes the quality of pretraining. The
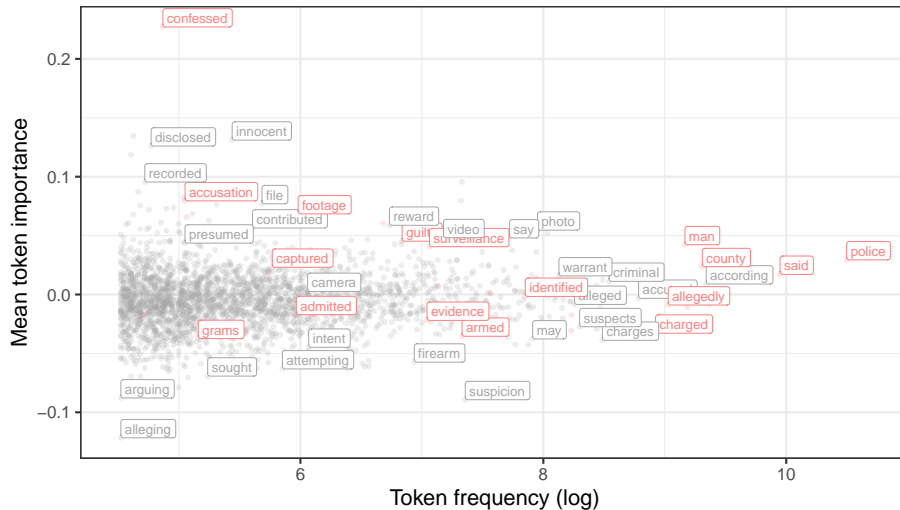
Figure 7: Mean model token importance by frequency. Words that received the most highlights overall (see Figure 3) are presented in red, other words in grey. In contrast to the highlighting data in Section 3.1, the token importance measure differentiates between tokens that increase (above 0 mean token importance) and decrease the predicted rating (below 0 mean token importance).

loss reduces up to 74%, suggesting that genre pretraining could significantly improve in-domain performance. We evaluate the end-to-end performance of the genre pretraining next.

For the core guilt-rating prediction tasks, we split the SuspectGuilt dataset into 85% training and 15% held-out test sets. We perform 5-fold cross validation and grid search on the training set. We then pick the best hyperparameters based on the best averaged loss of the 5-fold models, train our final model using the full training set for that fold, and report the final performance on the test set using the final model. We repeat the whole experiment with 20 different training-test splits to test the stability and significance of the performance. Additional details are given in Appendix B.2. We obtain a mean baseline by predicting everything as the mean values of the training set. We test the significance of *whether A is better than B* using the Wilcoxon signed-rank test (Wilcoxon, 1992).

5.2 Results and Discussion

Our results are summarized in Figure 6, which gives means and bootstrapped 95% confidence intervals. (Table 2 in our appendix gives the precise numerical values with standard deviations, and expands on the statistical analyses.)

The results suggest that *Author belief* is a harder task than *Reader perception*. This is aligned with the human results in Section 3.3.

In general, the mean-pooling models are sub-

stantially better than the CLS-based ones. Indeed, we fail to find evidence that BERT with the CLS token improves performance over the baseline ($p = 0.440$ for *Reader perception*; $p = 0.996$ for *Author belief*). Furthermore, when using both genre pretraining and token supervision, mean pooling is also significantly better than using the CLS token ($p = 0.001$ for *Reader perception*; $p = 0.022$ for *Author belief*).

Overall, a mean pooling model that makes use of genre pretraining as well as span-level supervision achieves the best performance. Span-level annotations are especially beneficial for the task of *Author belief* prediction, where this model significantly outperforms its closest competitors (e.g., when comparing against token supervision alone, $p = 0.022$). We thus conclude that both token-level supervision and genre pretraining provide important information for SuspectGuilt tasks.

6 Gradient-Based Token Importance

Although our models predict human guilt judgments well, the performance metrics don't tell us *how* they make predictions. Do they use information similarly to what we see in the human highlighting? Recent gradient-based methods for assessing feature importance in models like BERT (Sundararajan et al., 2017; Shrikumar et al., 2017) can help us answer this question.

Figure 7 presents one analysis of this form. We ran the Integrated Gradients method of Sundarara-

jan et al., as implemented in the PyTorch Captum library, on models which received genre pretraining but no highlighting supervision. The figure includes test-set runs averaged across 20 models with different random train–test splits. A positive score means that the token increases the predicted rating; a negative score corresponds to a decrease.

Like our highlighting data, the neural network's importance scores show the highest variance for words with low frequency. Words that received higher highlighting proportions for their frequency primarily affect the model predictions positively. In addition, we find that words that are more likely than random to be highlighted (as described in Section 3) are also significantly more likely to receive a higher token importance score in the model (Welch Two Sample t-test: $p < 0.01$). Beyond this, however, there is little correlation between the absolute attribution score for each word and its highlighting proportion ($r = 0.07$). While we can't rule out the possibility that this traces to the approximations introduced by Integrated Gradients, it seems likely that it helps explain why the span highlighting objective has a large impact on model predictions, as it is bringing in very different information than the model would otherwise attend to.

7   Conclusion

We introduced the SuspectGuilt corpus, which provides a basis for a quantitative study of how readers arrive at judgments of *Reader perception* and *Author belief*. We also showed that SuspectGuilt can be used to train predictive models on top of BERT parameters, and that these models are improved by genre-specific pretraining and supervision derived from token-level highlighting.

Understanding how news reporting affects reader judgments is a difficult task. The span-level highlighting in SuspectGuilt provides some insight into the factors at work here. We sought to match this with an introspective analysis of our predictive models using the gradient-based token importance method of Sundararajan et al. (2017). This yielded a very different picture from what we see in SuspectGuilt. Ultimately, this combination of annotations and model introspection might lead to new insights concerning how our models make decisions in this and other domains.

We also hope that this work paves the way to large-scale studies of how readers formulate judgments of guilt in crime reporting and encourages the development of systems that provide guidance on the presentation of these reports.

References

Jean-François Bonnefon and Gaëlle Villejoubert. 2006. Tactful or doubtful? expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, 17(9):747–751.

Aida Mostafazadeh Davani, Leigh Yeh, Mohammad Atari, Brendan Kennedy, Gwenyth Portillo Wightman, Elaine Gonzalez, Natalie Delong, Rhea Bhatia, Arineh Mirinjian, Xiang Ren, et al. 2019. Reporting the unreported: Event extraction for analyzing the local representation of hate crimes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5757–5761.

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.

Amanda M Durik, M Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 27(3):217–234.

Bonnie Erickson, E Allan Lind, Bruce C Johnson, and William M O'Barr. 1978. Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. *Journal of Experimental Social Psychology*, 14(3):266–279.

Caitlin M Fausey and Lera Boroditsky. 2010. Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review*, 17(5):644–650.

Scott Ferson, Jason O'Rawe, Andrei Antonenko, Jack Siegrist, James Mickley, Christian C Luhmann, Kari

Sentz, and Adam M Finkel. 2015. Natural language of uncertainty: numeric hedge words. *International Journal of Approximate Reasoning*, 57:19–39.

Kai von Fintel and Anthony S. Gillies. 2010. Must ... stay ... strong. *Natural Language Semantics*, 18(4):351–383.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of ACL*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english. *Proceedings of EMNLP-IJCNLP*.

Jakob D Jensen. 2008. Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. *Human communication research*, 34(3):347–369.

Kankawin Kowsrihawat, Peerapon Vateekul, and Prachya Boonkwan. 2018. Predicting judicial decisions of criminal cases from Thai Supreme Court using bi-directional GRU with attention mechanism. In *2018 5th Asian Conference on Defense Technology (ACDT)*, pages 50–55. IEEE.

Elisa Kreiss, Michael Franke, and Judith Degen. 2019. Uncertain evidence statements and guilt perception in iterative reproductions of crime stories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2020. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations*.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 480–489.

Marek Rei and Anders Søgaard. 2019. Jointly learning to label sentences and tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6916–6923.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org.

Mandy Simons. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056.

Matthew Stone. 1994. The reference argument of epistemic 'must'. Technical Report IRCS TR 97-06, Institute for Research in Cognitive Science, University of Pennsylvania.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Aviñante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 130–135. IEEE.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, Amherst, MA. GLSA Publications.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Appendices

A Data

2,818 annotators contributed to 3,463 submissions on Amazon's Mechanical Turk. The approximate time for completion was 15 minutes, and each participant was paid $2.50. We restricted participation to IP addresses within the US and an approval rate higher than 97%. Participants were asked to read 5 stories and respond to three questions about them (as described in Section 3.2). The full design of the trials is shown in Figure 8.

We excluded participants who indicated that they did the study incorrectly or were confused (544), whose self-reported native language was not English (71), who spent less than 3.5 minutes on the task (53), and who gave more then 2 out of 5 erroneous responses in the control questions (359). A response is considered erroneous when a clearly true or false question incorrectly received a slider value below or above 50 (the center of the scale) respectively. Additionally, we excluded 120 annotations because annotators had seen this story in a previous submission. Overall, we excluded 1,035 submissions and 120 annotations (15,405 annotations out of 51,945, resulting in 36,420 annotations).

A majority of annotators (89%) only participated once, which makes up 74% of all annotations. Only

14 annotators participated more than three times (0.7%).

The average age of annotators was 36 with a slightly higher proportion of male over female participants. The median time annotators spent on the study was 15.2 minutes, which is in-line with our original time estimates. Overall, annotators indicated that they enjoyed the study.

Annotators also had the option to indicate that the question cannot be applied to the news report. Overall, participants rarely used that option, but more so for the question about the *Author belief* (1.6%) than the *Reader perception* (10.5%) question. If several annotators agree that a question cannot be answered in the context of one particular story, it might be an indication that this story is not suitable for the corpus. We therefore decided to exclude stories where this box was selected more than 30% of the time with that particular question. Further inspection showed that this mainly affected summary news articles which addressed multiple stories and suspects and therefore the questions could not be uniquely attributed to one specific case.

B Experiments

B.1 Genre Pretraining

In this section, we describe the details of genre pretraining of BERT on our corpus. We set the maximum length to 400 tokens, with the tokens determined by the BERT tokenizer. This covers most of the instances in our corpus. We trained the model for 100K steps (roughly 30 epochs) using masked language modeling as described in (Devlin et al., 2019), with a mask probability of 0.15, a batch size of 128, and a learning rate of $5 \cdot 10^{-5}$. All experiments throughout this paper are based on PyTorch (Paszke et al., 2019) and Huggingface's Transformers (Wolf et al., 2019).

B.2 Predicting Guilt

In this section, we describe the hyperparameters used in our experiment.

For the basic models where there is no token supervision, we use the following hyperparameters

- Number of epochs: 5
- Warmup ratio: 10%
- Learning rate: 3E−5, 5E−5
- Random Seed: 0, 1
- Batch size: 16
- Checkpoints: 100 steps per checkpoint

1) Slider Rating

A Canton man accused of brandishing a handgun when his estranged wife showed up with another man to pick up their children is facing criminal charges, police said. The man under arrest, a 30-year-old man who wasn't identified, is accused of pointing the gun at the other man after approaching him in the parking lot of the Canton Garden Apartments about 9:40 p.m. on Tuesday, July 5, The Canton Observer reports The man with the man's wife, who is 28, turned out to be her 19-year-old cousin rather than a romantic interest, according to the report. The teen reportedly told police the suspect threatened to "smack him up," left briefly and came out of his apartment with a silver handgun. The teen took cover. The suspect's brother defused the situation, according to the report. Police confiscated the handgun and five rounds of ammunition and took the suspect into custody. » For more Canton police news, go to hometownlife.com.

How likely is it that the main suspect is / the main suspects are **guilty**?

very unlikely ————————————●———— very likely

☐ Doesn't apply.

SUBMIT RESPONSE

Optional: Is there something else you would like to add?

2) Highlights

A Canton man accused of brandishing a handgun when his estranged wife showed up with another man to pick up their children is facing criminal charges, police said. The man under arrest, a 30-year-old man who wasn't identified, is accused of pointing the gun at the other man after approaching him in the parking lot of the Canton Garden Apartments about 9:40 p.m. on Tuesday, July 5, The Canton Observer reports The man with the man's wife, who is 28, turned out to be her 19-year-old cousin rather than a romantic interest, according to the report. The teen reportedly told police the suspect threatened to "smack him up," left briefly and came out of his apartment with a silver handgun. The teen took cover. The suspect's brother defused the situation, according to the report. Police confiscated the handgun and five rounds of ammunition and took the suspect into custody. » For more Canton police news, go to hometownlife.com.

ERASE

How likely is it that the main suspect is / the main suspects are **guilty**?

Now please **highlight** in the text why you gave your response. If you're not happy with your selection, click on the ERASE button and start again.

NEXT

Optional: Is there something else you would like to add?

Figure 8: Participants rated a story on a continuous slider. After submitting, they highlighted the passages in the story that they considered to be most relevant for their assessment. At this point, they could not return to the previous screen to change the rating they gave.
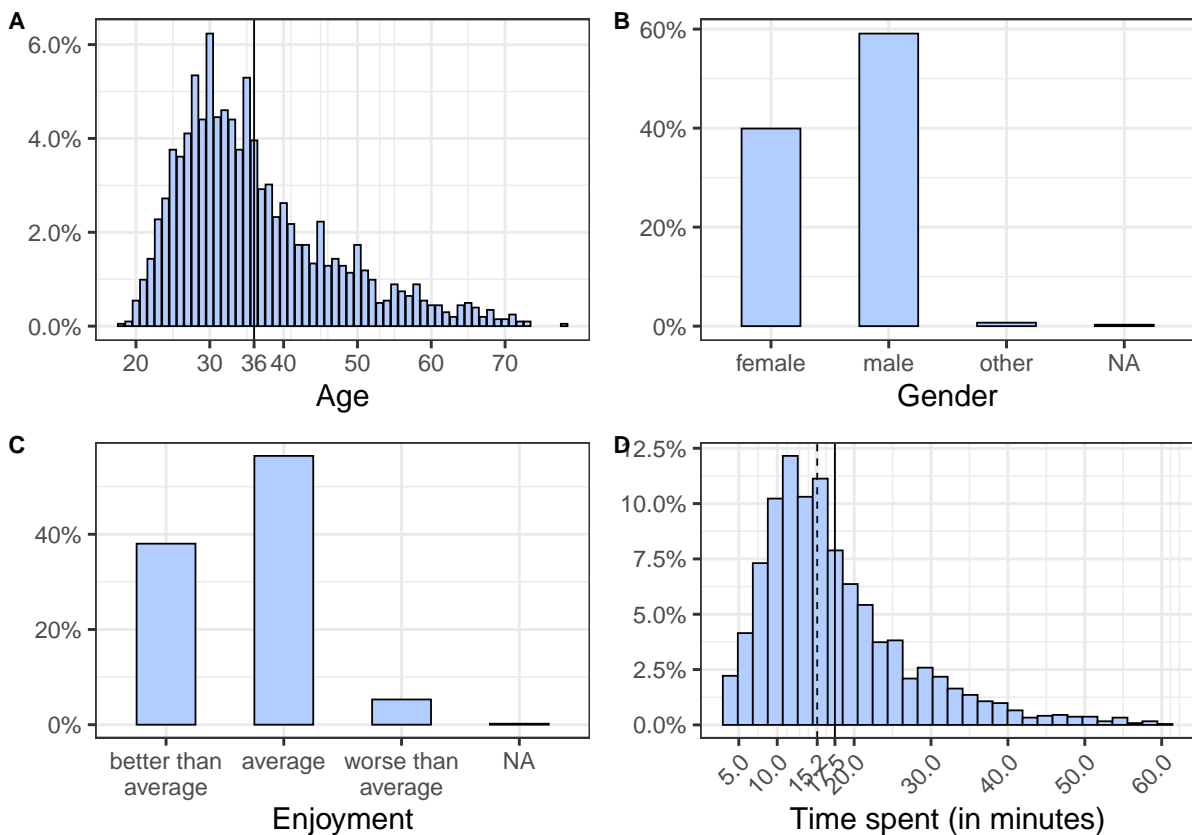

Figure 9: Participant demographics after exclusions.

	Reader perception Mean $_{\pm\text{std}}$	Author belief Mean $_{\pm\text{std}}$
Mean Baseline	$0.0119_{\pm 0.0009}$	$0.0121_{\pm 0.0010}$
BERT (CLS)	$0.0121_{\pm 0.0018}$	$0.0137_{\pm 0.0025}$
+ *pretraining*	$0.0104_{\pm 0.0013}$	$0.0114_{\pm 0.0007}$
+ *token supervision*	$0.0120_{\pm 0.0024}$	$0.0129_{\pm 0.0015}$
+ *pretraining + token supervision*	$0.0102_{\pm 0.0011}$	$0.0113_{\pm 0.0009}$
BERT (Mean)	$0.0106_{\pm 0.0013}$	$0.0113_{\pm 0.0012}$
+ *pretraining*	$0.0111_{\pm 0.0024}$	$0.0115_{\pm 0.0019}$
+ *token supervision*	$0.0096_{\pm 0.0009}$	$0.0113_{\pm 0.0011}$
+ *pretraining + token supervision*	$\mathbf{0.0095}_{\pm 0.0009}$	$\mathbf{0.0107}_{\pm 0.0011}$

Table 2: MSE for predicting guilt ratings for the *Reader perception* and *Author belief* questions. The models themselves are defined in Section 4. We report the mean and standard derivation values from 20 different runs. Bold denotes the best performance.

We also experimented with different number of epochs, batch sizes, and oversampling tail cases with different ratios in an initial small-scale study. We found that the current set of hyperparameters performs well in general. As adding more hyperparameter options is computationally intensive, we decided to use this set for our full-scale experiments.

When training the final model, we use the checkpoint whose corresponding steps are closet to 1.25 times the average number of steps of best performing checkpoints in the 5-fold cross validation.

For the models with token supervision, we use the same set of hyperparameters of no token supervision models except we only use one seed and add a hyperparameter of the loss ratio λ, with options of $[1, 2]$.

B.3 Numerical Results

Table 2 gives the corresponding numerical values for Figure 6. Whereas Figure 6 gives bootstrapped confidence intervals, here we given standard deviations to quantify the amount of variation seen across runs. Below are some additional details on these comparisons ('AB' = *Author belief*; 'RP' = *Reader perception*. Our statistical test here is the Wilcoxon signed-rank test.)

1. BERT with the CLS token does not improve performance compared to a simple mean baseline ($p = 0.449$ for RP and $p = 0.998$ for AB), while BERT with mean-pooling achieves better performance compared to the mean baseline ($p < 0.001$ for RP and $p = 0.004$ for AB).

2. The differences between using mean pooling and the CLS token are significant ($p = 0.003$ for RP and $p < 0.001$ for AB).

3. When using both the genre pretraining and the token supervision, mean pooling is significantly better than using the CLS token ($p = 0.001$ for RP and $p = 0.022$ for AB).

4. Overall, a mean pooling model that makes use of genre pretraining as well as span-level supervision achieves the best performance, significantly outperforming other models ($p < 0.001$ for RP and $p = 0.027$ for AB when comparing with the mean baseline; $p = 0.001$ for RP and $p = 0.020$ for AB with genre pretraining; and $p = 0.131$ for RP and $p = 0.022$ for AB with joint supervision).

5. Neither mean pooling models with genre pretraining ($p = 0.649$ for RP and $p = 0.464$ for AB) nor span-level supervision ($p = 0.001$ for RP and $p = 0.215$ for AB) alone can improve performance substantially in comparison to the mean baseline (only joint supervision for RP is significant).