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

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

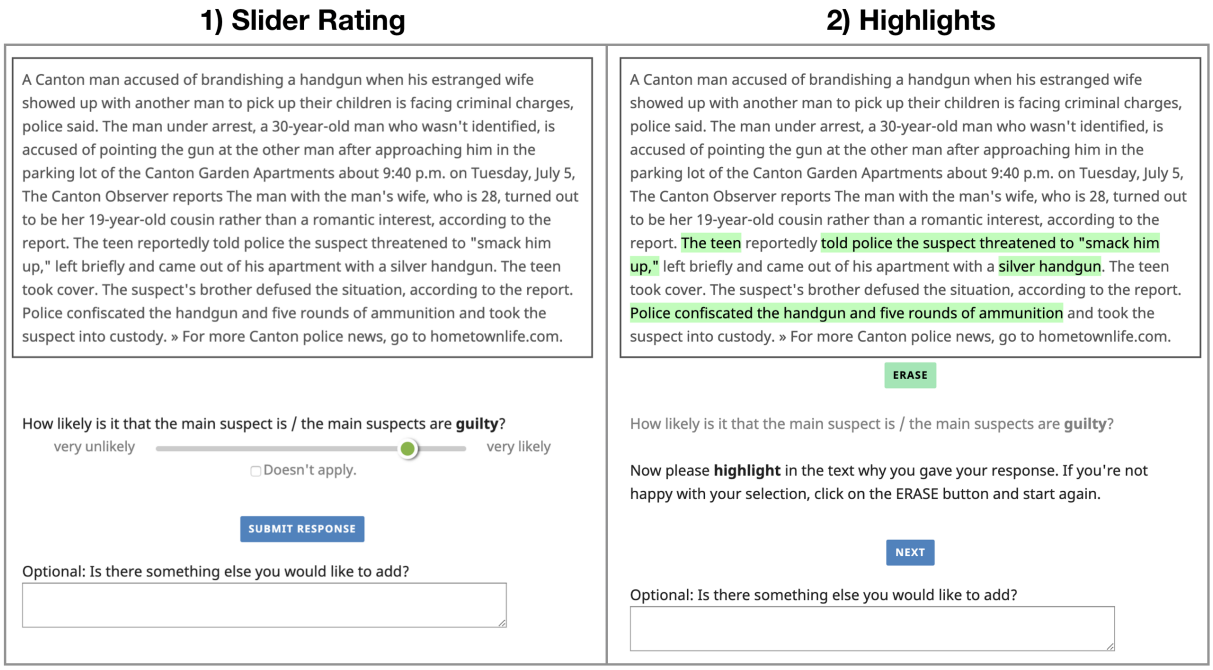

Figure 8: Participants rated a story on a continuous slider. After submitting, they highlighted the passages in the story that they considered to be most relevant for their assessment. At this point, they could not return to the previous screen to change the rating they gave.

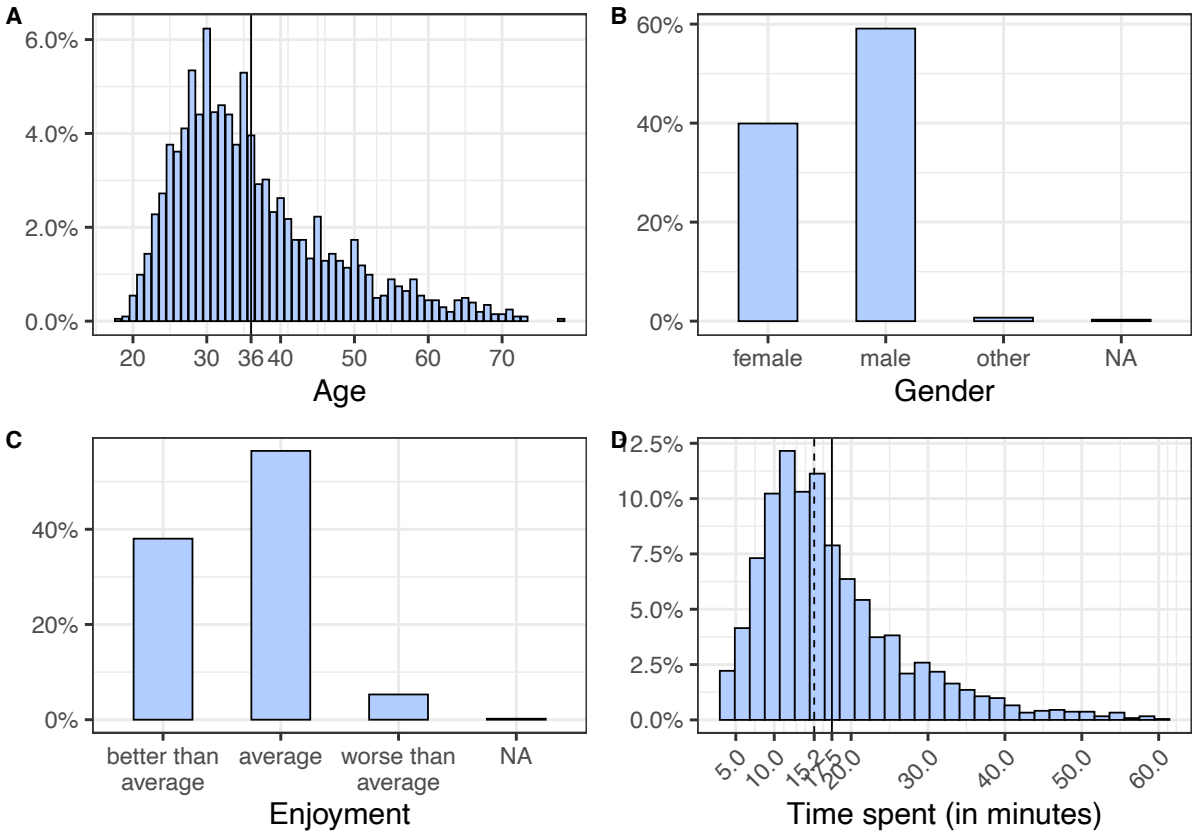

Figure 9: Participant demographics after exclusions.

	Reader perception Mean $\pm$std	*Author belief* Mean $\pm$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.

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).