Are Shortest Rationales the Best Explanations For Human Understanding?

Anonymous ACL submission

Abstract

Existing self-explaining models typically favor extracting the shortest rationales possible ("shortest yet coherent subset of input to predict the same label"), with the assumption that short rationales are more intuitive to humans, even though short rationales lead to lower accuracy. However, there is a lack of human studies on validating the effect of rationale length on human understanding. Is the shortest rationale indeed the most understandable for humans? To answer this question, we design a self-explaining model that can take controls on rationale length. Our model incorporates contextual information and supports flexibly extracting rationales at any target length. Through quantitative evaluation on model performance, we verify that our method LIMITEDINK outperforms existing self-explaining baselines on both end-task prediction and human-annotated rationale agreement. We use it to generate rationales at 5 length levels, and conduct user studies to understand how much rationale would be sufficient for humans to confidently make predictions. We show that while most prior work extracts 10%-30% of the text to be rationale, human accuracy tends to stabilize after seeing 40% of the full text. Our result suggests the need for more careful design of the best human rationales.

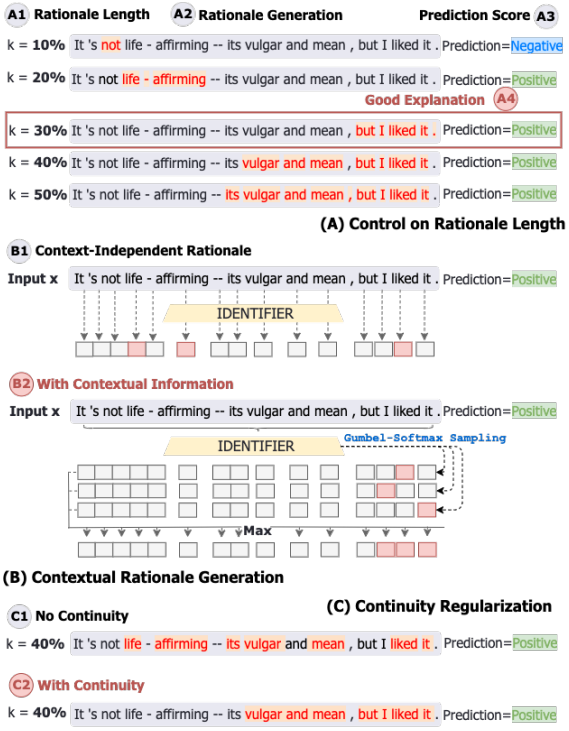


Figure 1: Our model design on rationale generation with length control. (A) control rationale generation with different lengths; (B) incorporating contextual information into rationale generation; (C) regularizing continuous rationale for human interpretability. Examples are from trained self-explaining models on SST dataset (Socher et al., 2013).

1 Introduction

As neural networks are achieving extraordinary prediction performance in dominating NLP tasks, it becomes increasingly important to explain why a model makes a specific prediction. Recent work starts to extract snippets of input text as the faithful rationale of prediction (Jain et al., 2020; Paranjape et al., 2020), with *rationale* defined as "shortest yet sufficient subset of input to predict the same label" (Lei et al., 2016; Bastings et al., 2019). The underneath assumption is two fold: (1) by retaining the label, we are extracting texts used by predictors (Jain et al., 2020); and (2) short rationales are more readable and intuitive for end users, and therefore are preferred for human understanding (Vafa et al., 2021). Importantly, prior work has knowingly traded off some amount of model performance in order to achieve shortest rationales. For example, when using less than 50% of text as rationales-for-predictions, Paranjape et al. (2020) achieved an accuracy of 84.0% (compared to 91.0% if using the full text). But existing work propose shortest rationales have better human interpretability by intuition rather than from empirical human studies (Vafa et al., 2021). Moreover, when the rationale is too short, the model has a much higher chance of missing the main point in the full text. In Figure 1(A), though the model is able to make the

correct positive prediction when using only 20% of the text, it relies on a particular adjective, "life-affirming", which is seemingly positive but does not reflect the author's sentiment. They may simply be confused when presented to end users.

In this work, we ask: *is shortest rationales really supportive of human understanding?* and examine the effects of rationale length on human understanding and performance. Our work includes two steps: First, we design and train a self-explaining model that allows for sparsity control. That is, the model can flexible extract rationales of a targeted length, such that we can compare user perceptions on a set of rationales with varying lengths. As shown in Figure 1, our model design consider three aspects: (A) controllability on rationale length, (B) being context-aware such to priorize certain amount of semantic information in the text, and, (C) extracting continuous text for readability. Through automated valuation on ERASER (DeYoung et al., 2019) datasets, we show that our model outperforms existing self-explaining baselines on both end-task prediction and rationale alignment with human ground annotations.

Using the rationales with different lengths generated from the model, we conduct human studies to evaluate human accuracy and confidence on predicting the document categories given only rationales. Our results show the best explanations for human understanding are largely not the shortest rationales. Given rationales with short length at 10%, human accuracy on predicting model class is worse than accuracy on the random baseline. Furthermore, while most prior work extracts 10%-30% of text to be rationale (Jain et al., 2020; Paranjape et al., 2020), human accuracy tend to stablize after seeing 40% of the full text. Our result sounds a cautionary note, and we encourage future work to more rigorously define or evaluate the typical assumption of "shorter rationales are easier to interpret" before trading off model accuracy for it.

2 LIMITEDINK

2.1 Self-Explaining Model Definition

We start by describing the typical self-explaining method (Lei et al., 2016). Consider a text classification dataset containing each document input as a tuple $(\mathbf{x}, y)$. Each input $\mathbf{x}$ includes n features (*e.g.,* sentences or tokens) as $\mathbf{x} = [x_1, x_2, ..., x_n]$, and y is the prediction. The model typically consists of a an *identifier* $\mathbf{idn}(\cdot)$ to derive a boolean mask $\mathbf{m} = \mathbf{idn}(\mathbf{x}) = [m_1, m_2, ..., m_n]$, where $m_i \in \{1, 0\}$ is a discrete binary variable. It then generates rationales $\mathbf{z}$ by $\mathbf{z} = \mathbf{m} \odot \mathbf{x}$, and further leverages a *classifier* $\mathbf{cls}(\cdot)$ to make prediction y based on the identified rationales as $y = \mathbf{cls}(\mathbf{z})$. The optimization objective is:

$$\min_{\theta_{\mathbf{idn}}, \theta_{\mathbf{cls}}} \underbrace{\mathbb{E}_{\mathbf{z} \sim \mathbf{idn}(\mathbf{x})} \mathcal{L}(\mathbf{cls}(\mathbf{z}), y)}_{\text{sufficient prediction}} + \underbrace{\lambda \Omega(\mathbf{m})}_{\text{regularization}} \tag{1}$$

where $\theta_{\mathbf{idn}}$ and $\theta_{\mathbf{cls}}$ are trainable parameters of *identifier* and *classifier*. $\Omega(\mathbf{m})$ is regularization function on mask and λ is the hyperparameter.

2.2 Generating Sparsity Controllable Rationales with Contextual Information

To enable length control on rationales, we add rationale length constraints on the self-explaining model. Assuming rationale length is k as prior knowledge, we enforce the generated boolean mask to sum up to k as $\mathbf{m} = \mathbf{idn}(\mathbf{x}, k), k = \sum_{i=1}^{n}(m_i)$. Existing self-explaining methods commonly solve this by assuming a fixed Bernoulli distribution over each input feature, thus generate each mask element m_i independently conditioned on each input feature x_i (see Fig 1(B1)) (Paranjape et al., 2020) . However, these methods potentially neglect the contextual input information. We leverage the *Concret Relaxation of Subset Sampling* technique (Chen et al., 2018) to incorporating contextual information into rationale generation process (see Fig 1(B2)), where we aim to select the top-k important features over all n features in input $\mathbf{x}$ during a weighted subset sampling process. To further empirically guarantee the precise rationale length control, we deploy a a *vector and sort* regularization on mask $\mathbf{m}$ (Fong et al., 2019). See more model details in Appendix A.1.

2.3 Regularizing Rationale Continuity

To enforce coherent rationale for human interpretability, we further employ the Fused Lasso to encourage continuity property (Jain et al., 2020; Bastings et al., 2019). The final regularization is:

$$\Omega(\mathbf{m}) = \lambda_1 \underbrace{\sum_{i=1}^{n} |m_i - m_{i-1}|}_{\text{Continuity}} + \lambda_2 \underbrace{\| \text{vecsort}\,(m) - \hat{m} \|}_{\text{Length Control}} \tag{2}$$

For BERT-based models using non-contiguous subword-based tokenizers (*e.g.,* WordPiece), we further assign the token's importance score as its sub-tokens' max score for rationale extraction during inference (see Fig 1(C)).

Method	Movies				BoolQ				Evidence Inference				MultiRC				FEVER			
	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1
Full Text	.90	-	-	-	.47	-	-	-	.48	-	-	-	.67	-	-	-	.89	-	-	-
Sparse-N	.79	.18	.36	.24	.43	.12	.10	.11	.39	.02	.14	.03	.60	.14	.35	.20	.83	.35	.49	.41
Sparse-C	.82	.17	.36	.23	.44	.15	.11	.13	.41	.03	.15	.05	.62	.15	.41	.22	.83	.35	.52	.42
Sparse IB	.84	.21	.42	.28	.46	**.17**	.15	.15	.43	.04	.21	.07	.62	**.20**	.33	.25	.85	**.37**	.50	**.43**
LIMITEDINK	**.91**	**.26**	**.88**	**.40**	**.62**	**.17**	**.67**	**.27**	**.50**	**.05**	**.44**	**.09**	**.68**	.16	**.90**	**.28**	**.90**	.28	**.67**	.39

Table 1: End-task predictive performance ("Task") and human annotated rationale agreement ("P"/"R"/"F1") on our LIMITEDINK and baselines. All results are on test sets and averaged across five random seeds.

3 Model Performance Evaluation

We next validate our model performance on end-task prediction and human annotation agreement.

3.1 Experimental Setup

We evaluate our method on five text classification datasets from ERASER benchmark. Our self-explaining models use use BERT-based modules. The *identifier* consists of a BERT-based model followed by two linear neural networks to encode representation and generate probability score for each feature. We further conduct the *concrete relaxation of subset sampling method* to convert the logit into binarized mask with predefined length. We empirically set five length levels from 10% to 50% with 10% interval. The *classifier* inputs the selected rationales to the BERT-based sequence classification module and outputs the final prediction.

We compare our method with four baselines. *Full Text* consists only *classifier* module with full text inputs. *Sparse-N* enforces shortest rationales by minimizing rationale mask length (Lei et al., 2016; Bastings et al., 2019). *Sparse-C* controls rationale length by penalizing the mask when its length is less than a threshold (Jain et al., 2020). *Sparse IB* enables length control by minimizing the KL-divergence between the generated mask with a prior distribution (Paranjape et al., 2020). See Appendix A.1 for more model and baseline details.

3.2 Evaluation Results

End-Task Prediction Performance. Following metrics in DeYoung et al. (2019), we report the weighted average F1 scores for classification tasks to evaluate end-task prediction performance. Choosing from the five self-explaining models with different rationale lengths, we report the optimal performance (varying depending on datasets and each baseline) as shown in Table 1. We observe our model consistently outperform the best self-explaining baselines with relative improvement from 5.88% (FEVER) to 34.78% (BoolQ). Further, our model can outperform full text inputs when only conditioning on extracted rationales, with relative improvement from 1.11% (Movies) to 31.91% (BoolQ). We further conduct ablation studies on each model components shown in Appendix A.2.

Human Annotated Rationale Agreement. We assess human plausibility automatically by evaluating the agreement between generated rationales and human annotations collected in ERASER benchmark (DeYoung et al., 2019). Also shown in Table 1, We report the Token-level F1 metric along with corresponding Precision (P) and Recall (R) scores. Results show our model improves the best baseline's Token F1 score with relative improvement from 12.00% (MultiRC) to 80.00% (BoolQ) on four datasets. However, our Token F1 score is lower than Sparse IB with 9.3% in FEVER dataset.

4 Human Studies

4.1 Experiments

Good explanations can justify the model predictions, humans should be able to predict the correct labels with high confidence given only generated rationales (Lertvittayakumjorn and Toni, 2019). Therefore, we design a human study to show humans with only model generated rationales, ask humans to predict the review label and provide a 5-point Likert scale confidence on their selection. In detail, conditioning on correct model predictions, we randomly sampled 100 reviews from Movie dataset (Zaidan and Eisner, 2008) and generated five rationales with lengths from 10% to 50% with an increment of 10%. In each task, we show humans five levels (10%-50%) of rationales one-by-one and asked their prediction with confidence. The five rationales's data index and order are all randomly sampled. In comparison, we design strict random baselines to contrast with the gain of just seeing more rationale length.

We use MTurk for the human study. We strictly control the worker group participation to make sure each worker only see a review once at a single length level, therefore eliminating learning effect. We collected 1150 assignments from 110 distinct workers. See more human study and user interface details in Appendix A.3.

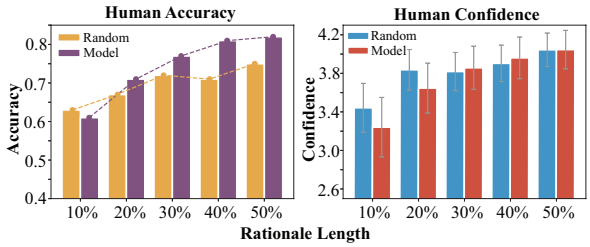


Figure 2: Humans overall accuracy and confidence performance on predicting rationale labels.

	Negative	Positive
	P / R / F1	P / R / F1
Model@10%	0.68 / 0.54 / / 0.60	0.66 / 0.58 / 0.62
Rand @10%	**0.68 / 0.53 / 0.60**	**0.63 / 0.71 / 0.67**
Model@20%	**0.75 / 0.61 / 0.67**	**0.72 / 0.77 / 0.74**
Rand @20%	0.69 / 0.58 / 0.63	0.67 / 0.74 / 0.70
Model@30%	**0.74 / 0.75 / 0.75**	**0.80 / 0.78 / 0.79**
Rand @30%	0.72 / 0.62 / 0.66	0.73 / 0.79 / 0.70
Model@40%	**0.84 / 0.76 / 0.80**	**0.78 / 0.85 / 0.81**
Rand @40%	0.79 / 0.63 / 0.70	0.65 / 0.79 / 0.72
Model@50%	**0.78 / 0.78 / 0.78**	**0.85 / 0.85 / 0.85**
Rand @50%	0.78 / 0.64 / 0.70	0.74 / 0.84 / 0.79

Table 2: Humans accuracy performance on predicting rationale labels for each class label in Movie dataset.

4.2 Results

We show human prediction accuracy and confidence results in Figure 2. We find that **best explanations for human understanding are largely not the shortest rationales** at 10% length level. In particular, when rationales are short at 10% length level, human accuracy on predicting model rationales are lower than random baseline (*i.e.,* 0.60 compared to 0.63), clearly indicating shortest rationales are not the best for human understanding. The different is statistically significant, with $p = .01$ with Student's t-test. The detailed human precision/recall/F1 scores are in Table 2.

Additionally, notice that the slope of our model's accuracy shows a consistently flatten as the rationale increases, whereas the random baseline does not display any apparent trend and obviously lower than our model at higher length levels (*e.g.*, 40%). We hypothesize that this means our model is (1) indeed learning to reveal useful rationales (rather than just randomly displaying meaningless text), and (2) the amount of information necessary for human understanding only start to saturate around 40% of the full text. This creates a clear contrast with prior work, where most studies extract 10%-30% of the text as the rationale on the same dataset (Jain et al., 2020; Paranjape et al., 2020).

5 Discussion and Limitation

While in Section 3 we validate our modeling approach through comparisons with baseline methods, in Section 4 we show that shortest rationales extract from our model are still not sufficient for human understanding. This contrast indicates that, extracting shortest text that still retain correct predictions — a standard definition for self-explanation models — may not necessarily support human understanding.

Of course, our finding is limited to the Movie Review dataset, and we predict that the optimal rationale length would be dataset dependent (e.g. short texts may even need a rationale of 80% to cover just five words). Still, our work sounds a cautionary note, and we encourage future work to more rigurously define or evaluate the typical assumption of "shorter rationales are easier to interpret" (Vafa et al., 2021; Bastings et al., 2019), before trading off model accuracy for it. One promising direction can be clearly define the optimal human interpretability in an measurable way, and then learn to adaptively select rationales with appropriate length.

6 Related Work

Current self-explaining models often enforce shortest yet sufficient rationales, with the assumption that short rationales are more intuitive to humans (Lei et al., 2016; Bastings et al., 2019). Paranjape et al. (2020) proposes an information bottleneck approach to enable the rationale length control. However current studies only assessed the methods with auto-metrics and did not evaluate human understanding on different rationale lengths. On the other hand, a line of studies measure the "human rationales alignment" (Paranjape et al., 2020), which compares how well the model generated rationales are agreeing with human grounded annotations (DeYoung et al., 2019). There are also studies involving human-in-the-loop to evaluate the explanations, such as asking humans to choose a better model Ribeiro et al. (2016). However, there is a lack of human evaluations on validating the effect of rationale length on human understanding.

7 Conclusion

To investigate if the shortest rationales are best understandable for humans, this work presents a self-explaining model that outperforms current baselines on both end-task prediction and human rationale alignment. we further use it to generate rationales for human studies to examine how rationale length can affect human understanding. Our results show shortest rationales are largely not the best for human understanding.

8 Ethical Considerations

This work investigates if the shortest rationales are best understandable for humans. We present a self-explaining model that incorporates contextual information to control rationale length. Here we examine the ethical considerations of this model by explicitly answering *what are the possible harms to users when the model is being used?*

When the model is used as intended and functions correctly, we note there are still potential risks. For example, when the rationales are incorrect, only showing rationales to humans might lead humans to misunderstand the model behavior and ignore some contents that are true cause of prediction or critical to them. Besides, if the model is trained from biased datasets, only showing rationales, although more interpretable for humans but hide much input information, can lead to biased judgement for humans. However, to mitigate these issues in real applications, we can keep "unimportant " features of input still present and especially highlight the rationales, so that humans can quickly capture the important features while able to comprehend the whole input context.

Furthermore, we are aware that some potential biases could be introduced (unexpectedly) to the users. For example, some informative words might be incorrectly removed or masked by the proposed methods and mislead users. To address the possible harms, we can *(i)* explicitly inform users the potential incorrectness of model behavior; and *(ii)* allow users to disagree or give feedback to the deployed method. Additionally, we set the MTurk workers to satisfy one qualification type as being "Adult", considering the case that instances in Movie dataset have sensitive information.

References

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR.

Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. *Proceedings of IEEE Conference on Machine Learning (ICML)*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. *Proceedings of International Conference on Learning Representations (ICLR)*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117. Association for Computational Linguistics.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. 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 5195–5205, Hong Kong, China. Association for Computational Linguistics.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "" why should i trust you?"" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*,

pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Keyon Vafa, Yuntian Deng, David M Blei, and Alexander M Rush. 2021. Rationales for sequential predictions. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

A Appendix

A.1 Model Details and Hyperparameters

A.1.1 Methodology Details

Concrete Relaxation of Subset Sampling Process. Given the output logits of *identifier*, we use Gumbel-softmax (Jang et al., 2017) to generate a concrete distribution as $\mathbf{c} = [c_1, ...c_n] \sim \text{Concrete}(\mathbf{idn}(\mathbf{x}))$, represented as a one-hot vector over n features where top important feature is 1. We then sample this process for k times in order to sample top-k important features, where we obtain k concrete distributions as $\{\mathbf{c}^1, ..., \mathbf{c}^k\}$. Next we define one n-dimensional random vector $\mathbf{m}$ to be element-wise maximum of these k concrete distributions along n features, denoted as $\mathbf{m} = \max_j\{\mathbf{c}_i^j\}_{i=n}^{j=k}$. Discarding the overlapping features to keep the rest, we then use $\mathbf{m}$ to as the k-hop vector to approximately select the top-k important features over document $\mathbf{x}$.

Vector and sort regularization. we deploy a a *vector and sort* regularization on mask $\mathbf{m}$ (Fong et al., 2019). , where we sort the output mask m in a increasing order and minimize the L_1 norm between m and a reference $\hat{m}$ consisting of $n - k$ zeros followed by k ones.

A.1.2 Model Training Details.

Training and inference: During training, we select the Adam optimizer with learning rate at 2e-5 with no decay. We set hyperparameters in Equation 5 and 2 as $\lambda = 1e - 4$, $v_1 = 0.5$ and $v_2 = 0.3$ and trained 6 epochs for all models. Furthermore, we trained LIMITEDINK on a set of sparsity levels as $k = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and chose models with optimal predictive performance.

A.1.3 Details of Self-Explaining Baselines

We compare our method with state-of-the-art self-explaining baseline models.

Sparse-N (Minimization Norm) This method learns short mask with minimal L_0 or L_1 norm (Lei et al., 2016; Bastings et al., 2019), which penalises for the total number of selected words in the explanation.

$$\min \mathbb{E}_{\mathbf{z}\sim\mathbf{idn}(\mathbf{x})}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda||m|| \quad (3)$$

Sparse-C (Controlled Norm Minimization) This method controls the mask sparsity through a tunable predefined sparsity level α (Chang et al.,

2020; Jain et al., 2020). The mask is penalized as below as long as the sparsity level α is passed.

$$\min \mathbb{E}_{\mathbf{z}\sim\mathbf{idn(x)}}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda \max(0, \frac{||\mathbf{m}||}{N} - \alpha) \tag{4}$$

where N is the input length and $||m||$ denotes mask penalty with L_1 norm.

Sparse IB (Controlled Sparsity with Information Bottleneck) This method introduces a prior probability of $\mathbf{z}$, which approximates the marginal $p(\mathbf{m})$ of mask distribution; and $p(\mathbf{m}|\mathbf{x})$ is the parametric posterior distribution over $\mathbf{m}$ conditioned on input $\mathbf{x}$ (Paranjape et al., 2020). They design the sparsity controll via the information loss term, which reduces the KL divergence between the posterior distribution $p(\mathbf{m}|\mathbf{x})$ that depends on $\mathbf{x}$ and a prior distribution $r(\mathbf{m})$ that is independent of $\mathbf{x}$.

$$\min \mathbb{E}_{\mathbf{z}\sim\mathbf{idn(x)}}\mathcal{L}(\mathbf{cls}(\mathbf{z}), y) + \lambda KL[p(\mathbf{m}|\mathbf{x}), r(\mathbf{m})] \tag{5}$$

A.2 Ablation Study on Model Components

We provide an ablation study on the Movie dataset to evaluate each loss term's influence on end-task prediction performance, including Precision, Recall, and F1 scores. The result is shown in Table 3.

Setups	End-Task Prediction		
	Precision	Recall	F1
No Sufficiency	0.25	0.50	0.34
No Continuity	0.82	0.81	0.81
No Sparsity	0.80	0.79	0.79
No Contextual	0.83	0.83	0.83
Our Model	**0.92**	**0.91**	**0.91**

Table 3: Ablation study of each module in our model on Movie dataset.

A.3 Additional Details of Human Evaluation

A.3.1 Additional Details of Human Study

Random Baseline Design. We design the random baseline to be also continuous, keeping same total tokens and averaged number of chunks as our model generated rationales on each length level. Specifically, given the sparsity level k, we get the count of total tokens in rationale as #tokens $=$ #input_length $* k$; we compute the average spans count over dataset generated by our model m; we generate m random integers with fixed sum at k, meaning dividing the baseline review randomly into m spans with length of these values; Finally, we randomly chose the start position of these m spans for rationales.

Control Experiment Design. To strictly control the experiments, we grouped 5 reviews into one batch and obtain 20 batches in total. For each batch, we created 10 tasks (webpages) and assign 10 worker groups to conduct the human study. We used costum MTurk qualifications to strictly control worker participants, so that workers who joined one group could not view tasks from other groups. provide detailed worker group control design in Figure 3(A).

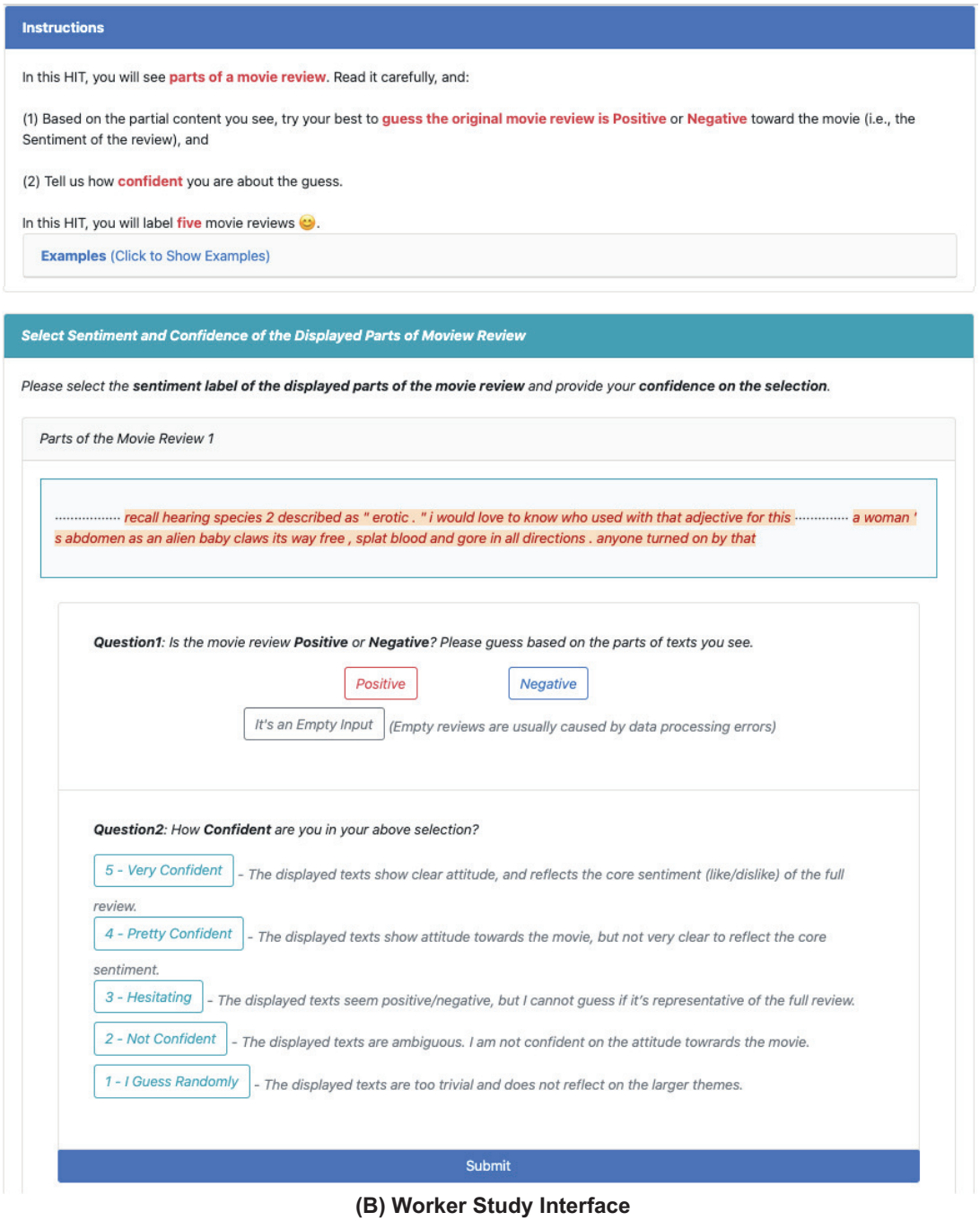
Amazon MTurk Study Statistics. We present each task to 7 MTurk workers. In first stage – worker recruiting stage – we recruited 200 crowd workers where each worker finished one simple assignment. We conduct our human study in the second stage with the recruited 200 workers. There are 110 out of the distinct workers participated and finished 1150 assignments in our study. We compensate workers at a rate of \$0.50 per assignment in worker recruiting and \$0.20 per assignment in task evaluation. Our assignment response rate is 84.38% in total.

A.3.2 Human Evaluation User Interface

We provide our designed user interfaces used in the human study. Specifically, we show the interface of human study panel in Figure 3 (B). We also provide the detailed instructions for workers to understand our task, the instruction inteface is shown in Figure 4.

	Review1	Review2	Review3	Review4	Review5
Worker Group 1	Our@10%	Our@20%	Our@30%	Our@40%	Our@50%
Worker Group 2	Our@20%	Our@30%	Our@40%	Our@50%	Our@10%
Worker Group 3	Our@30%	Our@40%	Our@50%	Our@10%	Our@20%
Worker Group 4	Our@40%	Our@50%	Our@10%	Our@20%	Our@30%
Worker Group 5	Our@50%	Our@10%	Our@20%	Our@30%	Our@40%
Worker Group 6	Random@10%	Random@20%	Random@30%	Random@40%	Random@50%
Worker Group 7	Random@20%	Random@30%	Random@40%	Random@50%	Random@10%
Worker Group 8	Random@30%	Random@40%	Random@50%	Random@10%	Random@20%
Worker Group 9	Random@40%	Random@50%	Random@10%	Random@20%	Random@30%
Worker Group 10	Random@50%	Random@10%	Random@20%	Random@30%	Random@40%

(A) Worker Group Assignment



(B) Worker Study Interface

Figure 3: (A) The design of worker group assignment in our human study. (B) User Interface of human study.

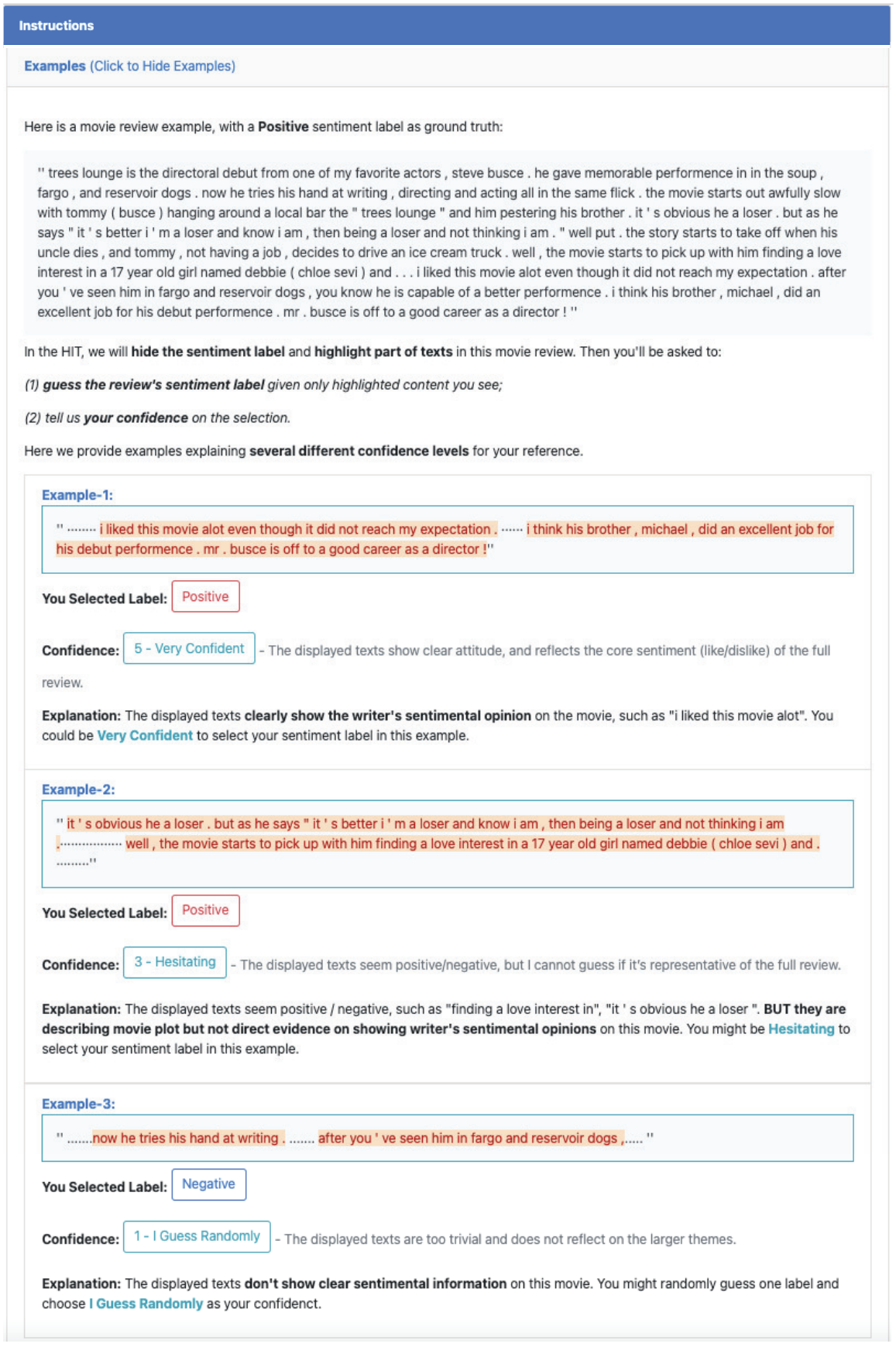


Figure 4: User Interface of the instruction in the human study.