Attention Highlights Help Humans Predict Model Behavior

Anonymous ACL submission

Abstract

Although numerous studies have investigated whether or not attention can be used by *researchers* as a tool of interpretability for understanding their models, a consensus has yet to be reached. This study aims to examine the attention mechanism practicality by testing whether attention can help the *end-users* of such models to predict their behaviour by comparing their performance with and without access to attention highlights. We divided human evaluators into two groups—one with access to attention highlights and another without—to assess the performance differences between the two groups in terms of decision-making accuracy and response time. Our results showed that including attention highlights significantly improved decision-making accuracy for humans to predict extractive question-answering model output, with a notable difference in F_1 scores between the two groups. However, the time taken to predict model response was not significantly affected, suggesting that attention highlights did not speed up decision-making.

1 Introduction

The mechanism of attention in neural networks builds on the idea that a more informative representation at any decision-making step in the neural network can be calculated by averaging current representations from previous rounds of propagation in the network, weighting them by learned attention weights. Since the introduction of the attention mechanism in deep learning models (Bahdanau et al., 2015), especially since the Transformer architecture (Vaswani et al., 2017) became prevalent in natural language processing (NLP), much work has followed using the so-called *attention weights* as an interpretability tool for gaining a better understanding of the underlying decision-making mechanisms of NLP models (Li et al., 2016; Mullenbach et al., 2018; Abnar and Zuidema, 2020).

The great popularity of using attention as an interpretability method has led to a heated debate about whether researchers can use it as a faithful explanation for their models' decisions or not (Serrano and Smith, 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Sood et al., 2020; Ethayarajh and Jurafsky, 2021).[1] While many researchers have extensively investigated the use of attention mechanisms to reinforce findings about their models' decision-making process, the extent of attention's practical usefulness for end-users of such models remains to be determined. More precisely, this paper aims to answer the question **can attention weights assist humans in making more informed decisions?** We specifically address cases where the network's task aligns directly with the task humans need to perform, with a particular emphasis on the question-answering task.

Motivated by the safety implications of AI and discussions around evaluation and monitoring, we ideally want humans to be able to learn to predict model behaviour to prevent undesired outcomes before deployment. We seek to understand how attention weights can contribute to human-in-the-loop decision-making and, in turn, promote the development of safe and human-centred AI systems that leverage these mechanisms to enhance user understanding and aid in complex tasks like question answering. At the foundation of our experiments, we recruit human annotators and present them with a question and context with the aim of studying the mean difference between those who have access to attention weights in the form of highlights and the group that does not. Our findings demonstrate that annotators in the treatment group, equipped with attention highlights, attain an average F1 score more than 15% higher than their counterparts in the control group. This suggests that, despite the

[1] For a more comprehensive and detailed discussion on this debate, interested readers are encouraged to refer to the survey by Bibal et al. (2022).

ongoing debate among researchers regarding the validity of attention mechanisms as interpretability tools, the practical understanding of models' behaviour by end-users can substantially benefit from their incorporation.

2 Related Work

Connections between Human and Machine Attention Another closely related line of research is the examination of relations between models' attention and the human gaze. For example, Hollenstein et al. (2021) demonstrated that transformer models implicitly represent the relative importance of language, akin to human cognitive processing mechanisms. Morger et al. (2022) provided evidence that the first-layer attention and attention flow strongly correlate with human eye-tracking data in German, Dutch, English, and Russian datasets.

Human Attention as Input for NLP Models While the usefulness of models' attention for humans has hardly been explored, a vast body of work examines the opposite direction. Augmenting NLP models with human gaze data, which can be considered as "human attention", has been shown to improve performance across many tasks, such as named entity recognition (Hollenstein and Zhang, 2019), sentiment and sarcasm classification (Mishra et al., 2017), and grammatical error detection (Barrett et al., 2018), to name a few. According to Malmaud et al. (2020), the task we experiment with, the reading comprehension task, is well-fitted to establish a connection between human eye movement data and NLP modeling. This suitability arises from the significant alignment observed between reading times and the relevance of specific text segments in formulating answers to questions. Hahn and Keller (2023) provide additional evidence to support their assertion by demonstrating an evident rise in reading times when the correct answer is a named entity in a question-answering task.

We view our work as a complementary perspective to the aforementioned studies, as our findings demonstrate how machine attention can assist humans in understanding NLP models' decisions. This connection holds the potential to enhance the interpretability of NLP models for end-users.

3 Preliminaries

Extractive Question Answering In this work, we focus on extractive open-domain QA (also referred to as reading comprehension; RC), which is defined as the task of identifying a span in a given textual context that best answers a user question on a broad set of topics or domains. More specifically, we define three components: *Context* (C) is the textual information or passage containing verbatim the answer to the question. It is represented as $C = (c_1, \ldots, c_N)$, where c_i denotes the ith token in the context, and N is the total number of tokens. *Question* (Q) is the question for which an answer is sought from the context. It is represented as $Q = (q_1, \ldots, q_M)$, where q_i denotes the ith word or token in the question, and M is the total number of words or tokens. *Answer* (A) is a span of contiguous tokens in context C denoted as $A = (c_i, \ldots, c_j) \in C$ where $1 \leq i \leq j \leq N$. Note that while the context vocabulary V_C and question vocabulary V_Q may not be identical, there exists a relevant sub-vocabulary $V_R \subseteq V_Q \cap V_C$ that overlaps between the context and question that enables the model to match related semantics.

Attention Aggregation Various methods exist for aggregating attention. Central to our thought process is the understanding that attention operates as an additive (linear combination) mechanism (Elhage et al., 2022). Therefore, merely focusing on the attention weight in the layer preceding the output layer might not offer substantial information. Conversely, previous research indicates that attention aggregation and flow can deliver a more comprehensive strategy for examining the true significance of individual tokens in the output (Abnar and Zuidema, 2020). In this work, we use attention aggregation. Formally, for a given model with L layers and H heads, each task with an input sequence of length T, the attention weights for each head $h \in \{1, \ldots, H\}$ in each layer $\ell \in \{1, \ldots, L\}$ can be represented as a $T \times T$ matrix, $\Theta^{(l,h)}$, where the element of each matrix at position (i, j) represents the attention weight of the ith token attending to the jth token for the relevant head and layer. We aggregate the attention weights across all heads and layers into $\Theta \in \mathbb{R}^{T \times T}$, retaining the top-$k$ tokens with the highest attention weight:

$$\Theta = \frac{1}{LH} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \Theta^{(\ell,h)}. \quad (172)$$

4 Experimental Setup

Our experimental setup is quite straightforward. First, we train a model for solving the task of ex-

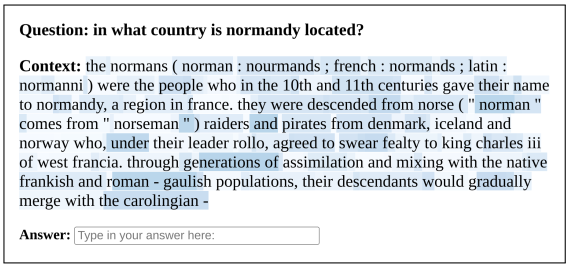

Figure 1: Example of a question presented to human annotators with the context highlighted according to the model attention weights.

tractive question-answering. We then aggregate the model's attention weights for its correct predictions and incorporate them into the corresponding subset of contexts to enrich them (see Figure 1). We followed by asking two distinct groups of human annotators to answer the same set of questions, where the treatment group used the enriched contexts, and the control group used the original ones. Our complete pipeline is presented in Figure 2.

Data We use a subset of contexts (C) and their corresponding questions (Q) from SQuAD 2.0 (Rajpurkar et al., 2018). To foster an assessment based on contextual understanding rather than individual world knowledge, we chose questions spanning six varied sub-domains: Normans, Computational Complexity Theory, Southern California, Sky (United Kingdom), Viktoria (Australia), Huguenot, and Steam Engine. Each domain consists of 14 questions, and seven annotators from the control and treatment groups answered each question. We provided human evaluators with questions that our *model answered correctly*. This approach allows us to ask annotators to predict the correct answer, making the task more straightforward. Our initial attempts to have annotators predict the model's answer received negative feedback, with annotators finding it unclear and counterintuitive. For further discussion of this issue, see Section 6.

Model We employ DistilBERT (Sanh et al., 2019) as our backbone model, a more inference-efficient version of BERT (Devlin et al., 2018). Our emphasis on DistilBERT stems from its efficiency, enabling deployment on end-users' devices for a more realistic setup in NLP applications. While large language models (LLMs) have gained much attention in recent years, we chose to use a model from the BERT family, as such models are currently state-of-the-art in question-answering.[2] We employ a variant of model-agnostic meta-learning (MAML; Finn et al. 2017), specifically fine-tuning the classifier (Raghu et al., 2019). We use meta-learning due to its unparalleled adaptability and rapid learning across diverse tasks—qualities absent in traditional task-specific models. For more details about the model, see Appendix A.

Human Evaluation We used Amazon Mechanical Turk (AMTurk)[3] to recruit participants and employed a custom-built interface for the experiment. We chose participants based on the ethics rules of our institutions[4]. The study is designed to have a control and treatment group, both of which receive the same inputs that the model receives (i.e., question and context), in addition to these inputs, the treatment group also receives information from the attention weights of the model displayed as word highlights on the context that mark the top-k attention scored words. In order to analyze the impact of attention on guiding participants' responses, the participants are not presented with the model predictions. For more details about our evaluation process, please refer to Appendix B.

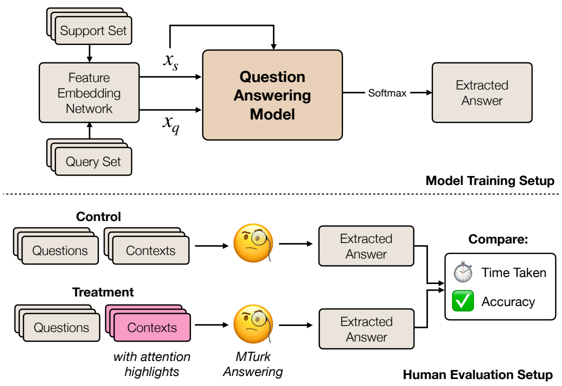

Figure 2: The overall architecture of our method: We first train a question-answering system using few-shot learning (for more details, see Appendix A). We then extract the attention weights across all heads and layers (represented in red) and reconstruct the question and the context. We show the human evaluators (treatment group) these questions alongside the context and record. The difference between the two groups is that treatment receives attention highlights.

[2]The Stanford leaderboard for SQuAD2.0 demonstrates the encoder-only models clear advantage - https://rajpurkar.github.io/SQuAD-explorer/.

[3]Refer to https://www.mturk.com/worker/help for information about AMTurk.

[4]Not referenced in this manuscript in order to preserve anonymity. We will reference the ethics rules in the camera-ready version.

Category	Attention Highlight			No Attention Highlight			
	Precision	Recall	F_1 Score	Precision	Recall	F_1 Score	Δ F_1 **Score**
Normans	82.9%	87.3%	84.4%	75.4%	78.9%	76.4%	↑8.0%
Computational Complexity Theory	80.4%	81.5%	79.7%	61.2%	65.3%	60.4%	↑19.3%
Southern California	73.0%	88.7%	73.2%	60.0%	68.6%	52.1%	↑21.2%
Sky (United Kingdom)	83.3%	85.1%	83.6%	81.7%	67.8%	66.3%	↑17.2%
Victoria (Australia)	86.2%	70.7%	76.3%	78.0%	51.5%	58.1%	↑18.2%
Huguenot	53.6%	82.7%	61.3%	43.5%	57.3%	45.0%	↑16.3%
Steam Engine	86.2%	46.8%	59.7%	81.2%	44.5%	53.4%	↑6.3%
Mean	77.9%	77.5%	74.0%	68.7%	62.0%	58.8%	↑15.2%

Table 1: Precision, recall and F1 metrics for each category of question type.

Extract QA	w/ Attention	w/o Attention
Time	22.17s	19.67s
Exact Match	44.0%	36.9%
F_1	74.0%	58.8%

Table 2: The performance results for the test set of SQuAD2.0, evaluated using exact-match.

5 Results

Table 1 describes the performance of the human performance with and without presenting them with the attention weights. Our results demonstrate a statistically significant increase in human performance when attention highlights are displayed, with the *Southern California* category having the largest impact. On average, though, it takes around 2 seconds more to process the documents with the attention weights (Table 2). In terms of mean time per group, the average processing time for the group with attention highlights was 22.17s, while the average time for the group without attention highlights was 19.67s. Moreover, we observed that the group without attention tended to skim through large portions of the text and quickly paste them into the answer box, while the group with attention focused on identifying relevant tokens carefully before pasting them into the answer box, as depicted in Figure 3.

Our analysis indicates that attention highlights do have a significant impact on the accuracy of human decision-making, as evidenced by the improved performance of the group which uses attention in terms of F_1 scores. When comparing the two average times it takes to complete the task for both groups (with and without attention), we observe no statistically significant difference according to a t-test with a significance level of $\alpha = 0.01$, indicating that attention highlights might not affect the speed of human decision-making. Overall, our results suggest that displaying attention highlights aids humans in locating relevant information in a question-answering task.

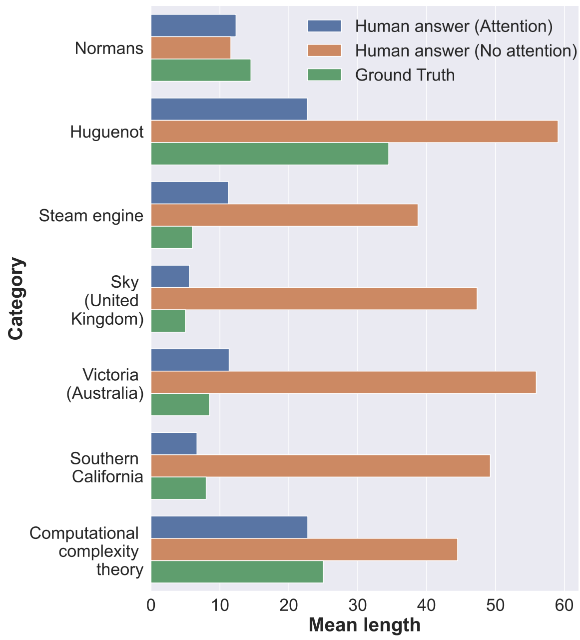

Figure 3: Illustrating the influence of highlight on answer length. It is evident that subjects, when not provided with a highlight, have a tendency to copy and paste a significant subsection of the context.

6 Conclusion

In summary, our direct approach, spotlighting top-k aggregated tokens determined by attention weights, adeptly points out potential answer locations within a document. The proven effectiveness of attention weights as a valuable tool for human interpretation of the decision-making process in Transformer models underscores the importance of this methodology. Subsequent research endeavors may delve into broader implications and applications, exploring the use of attention highlights across diverse scenarios. This exploration presents opportunities to elevate interpretability in a variety of domains.

Limitations

Constrained by limited resources, our study focuses on evaluating a single task using one dataset and a specific method to aggregate attention weights across a transformer model's multiple heads and layers. Consequently, the generalizability of our results to diverse settings is constrained. Further investigation is imperative to ascertain the applicability of our findings across various datasets, models, and aggregation methods. Moreover, effectively controlling variables such as participants' familiarity with the material being tested or the attention devoted to each question poses a substantial challenge. This paper examines how attention highlights improve end-users' ability to predict model behaviour, which can be crucial when models produce wrong predictions. Unfortunately, our attempts to ask annotators to predict the model behaviour, i.e., to predict its answer regardless of correctness, have proven to be perplexing for crowd-workers. We hence cannot guarantee our findings will generalized for cases where the model's predictions are wrong. In future work, we plan to expend our efforts on cases where the model answers are wrong, which will require a more customized experimental setting.

Ethics Statement

Human workers were informed of the intended use of the provided annotations and complied with the terms and conditions of the experiment, as specified by Amazon Mechanical Turk.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning*, pages 302–312.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, And Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. 2022. Softmax linear units. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/solu/index.html.

Kawin Ethayarajh and Dan Jurafsky. 2021. Attention flows are shapley value explanations. *arXiv preprint arXiv:2105.14652*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks.

Michael Hahn and Frank Keller. 2023. Modeling task effects in human reading with neural network-based attention. *Cognition*, 230:105289.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. *arXiv preprint arXiv:2104.05433*.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving ner with eye movement information. *arXiv preprint arXiv:1902.10068*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020. Bridging information-seeking human gaze and machine reading comprehension. *arXiv preprint arXiv:2009.14780*.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Felix Morger, Stephanie Brandl, Lisa Beinborn, and Nora Hollenstein. 2022. A cross-lingual comparison of human and model relative word importance. In *Proceedings of the 2022 CLASP Conference on (Dis) embodiment*, pages 11–23.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. 2019. Rapid learning or feature reuse? towards understanding the effectiveness of maml.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. *arXiv preprint arXiv:2010.06396*.

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.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

A Few-shot QA

As in the paper, we adopt a variant of model-agnostic meta-learning (MAML) (Finn et al., 2017) for few-shot question answering. Specifically, we only fine-tune the classifier while the feature extractor parameters are shared across tasks. We define a question answering task τ consisting of: Contexts $C_\tau = c_1, \dots, c_N$, Questions $Q_\tau = q_1, \dots, q_M$, Answer spans $A_\tau = (c_i, \dots, c_j) \in C_\tau$ The meta-learner model f_θ maps questions to predicted answer spans: $f_\theta : Q_\tau \to A_\tau$ In the N-way K-shot learning formulation, each task τ has a support set $\mathcal{S_T}$ with N classes and K labeled examples per class. The model is evaluated on query examples $\mathcal{Q_T}$ from the same classes. We group tasks into categories $\mathcal{C}$ by mapping related topics into each category (e.g. Sport, Education). During meta-training, f_θ is fine-tuned on the support sets $\mathcal{S_T}$ and evaluated on the query sets $\mathcal{Q_T}$ for each task τ to learn across tasks and categories. The meta-learning objective is: $\min_\theta \sum_{\tau \sim p(\tau)} \mathcal{L}(f_\theta^{\mathcal{S_T}}; \mathcal{Q_T})$ where $p(\tau)$ is the distribution over tasks, $f_\theta^{\mathcal{S_T}}$ is the model fine-tuned on support set $\mathcal{S_T}$, and $\mathcal{L}$ is the loss on the query set.

B AMTurk Details

We used Amazon Mechanical Turk (AMTurk)[5] to recruit participants and employed a custom-built interface for the experiment. We chose participants based on the ethics rules of our institutions[6]. Participant selection was limited to those in a single English-speaking country who had at least a university degree, ensuring a higher level of expertise in the subject matter. We followed ethical guidelines when compensating participants for their time and effort in providing valid responses, setting a task reward of $0.15 per assignment (calculated by hourly living wage divided by the approximate minimal time it takes to complete the assignment).

Before starting the main experiment, participants were given several practice questions in the same format as the actual questions to become accustomed to the setup and interface and were also shown the guidelines in Figure 4. We create a Human Intelligence Task (HIT) in AMTurk with the title "Answer this simple question, given a short

[5]Refer to `https://www.mturk.com/worker/help` for information about AMTurk.

[6]Not referenced in this manuscript in order to preserve anonymity. We will reference the ethics rules in the camera-ready version.

Step 1: Thank you for choosing to participate in this experiment. In this task we ask you to answer questions based on the context which is provided, you should be able to extract the answer from within the context. The experiment will start with a few practice questions, once you have completed the practice questions, you will be directed to answer the actual questions.

Step 2: You are presented with a Question and a Context. Read the two carefully and answer the Question in the "Answer" box. To answer the following question, you can:

1. You can copy-paste part of the context to answer the question, or just type the answer in your own words;
2. Type NOANSWER if the context doesn't contain the answer;
3. Type UNCLEAR if the question is unclear, in other words, you cannot understand what is being asked.

Answers are case-insensitive, in other words, capitalisation doesn't matter.

Figure 4: Guidelines provided to Amazon Mechanical Turk workers.

context snippet." and the description "Read this question with corresponding context, and write the answer (if it exists in the context)." The HIT has the keywords "text, quick, labeling" and a maximum of seven assignments are allowed per HIT. The HIT has a lifetime of 1 day and an assignment duration of 10 minutes. The auto-approval delay is set to 4 hours. The HIT has several qualification requirements: the worker's percentage of approved HITs must be greater than or equal to 98%, they must have at least 500 approved HITs, and they must have opted-in to view adult content. The custom interface is shown in figure 1.