Semantic Categorization of Social Knowledge for Commonsense Question Answering

Gengyu Wang[1] **Xiaochen Hou**[2]
Diyi Yang[3] **Kathleen McKeown**[1] **Jing Huang**[2]
[1]Columbia University, New York, NY
[2]JD AI Research, Mountain View, CA
[3]Georgia Institute of Technology, Atlanta, GA
gengyu.wang@columbia.edu, houxiaochen1994@hotmail.com
diyi.yang@cc.gatech.edu, kathy@cs.columbia.edu, jing.huang@jd.com

Abstract

Large pre-trained language models (PLMs) have led to great success on various commonsense question answering (QA) tasks in an end-to-end fashion. However, little attention has been paid to what commonsense knowledge is needed to deeply characterize these QA tasks. In this work, we proposed to categorize the semantics needed for these tasks using the SocialIQA as an example. Building upon our labeled social knowledge categories dataset on top of SocialIQA, we further train neural QA models to incorporate such social knowledge categories and relation information from a knowledge base. Unlike previous work, we observe our models with semantic categorizations of social knowledge can achieve comparable performance with a relatively simple model and smaller size compared to other complex approaches.

1 Introduction

Recently, large pre-trained language models (PLMs) (Devlin et al., 2019; Raffel et al., 2019; Liu et al., 2019) have been widely used on various commonsense QA tasks such as CommonsenseQA (Malaviya et al., 2020), SocialIQA (Sap et al., 2019b), and Mostafazadeh et al. (2016); Huang et al. (2019); Boratko et al. (2020); Levesque et al. (2012); Roemmele et al. (2011). One line of work (Khashabi et al., 2020) improved the performances of these QA tasks by aggregating more QA data and using even bigger PLM T5 (Raffel et al., 2019). Other line of work tried to supplement the question context with retrieval of related knowledge from external knowledge bases (KB), or re-trained PLMs under the guidance of KBs (Shen et al., 2020; Shwartz et al., 2020; Mitra et al., 2019; Ji et al., 2020a,b).

However, very little past research has paid attention to the specific question/context knowledge types that are needed for these commonsense QA tasks. Therefore, in this paper, we go deeper into the QA task context and take a closer look at the semantics on what additional information can be inferred from the given question-answer context in order to answer a question. Using SocialIQA as an example, we propose to add two new context types (See Figure 1) into the neural QA model: one on question relation type derived from ATOMIC (which was used to create SocialIQA), and another knowledge category type from our own constructed social knowledge taxonomy. While the question

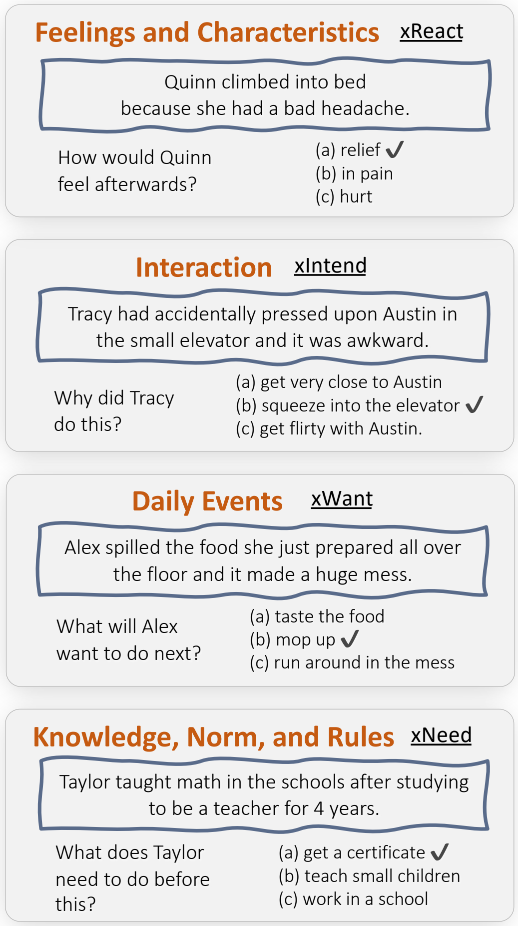

Figure 1: SocialIQA Examples for Social Knowledge Categories and **Question Relation Type**

Proceedings of the 2nd Workshop on Simple and Efficient Natural Language Processing, pages 79–85
November 10, 2021.

relation type derived from ATOMIC is restricted on ATOMIC related datasets, our constructed social knowledge category type has the potential be generally applied to other social knowledge related tasks.

To fully utilize these two new types of context information, we adopt a simple yet effective way of integrating this information to help in the neural QA model. Specifically, we concatenate each QA pair with its assigned question relation type or its social knowledge category as the input to a PLM (say RoBERTa (Liu et al., 2019)), and fine tune the RoBERTa model for the SocialIQA task. Our experimental results show that this simple and interpretable method not only outperforms the RoBERTa baseline model, but also achieves comparable performances as that of previous work, which adopted much more complex models to encode external knowledge or re-train large language models.

In terms of creating efficient and sustainable models for QA tasks, our work illustrates the importance of deep understanding of what knowledge is required for the specific commonsense tasks. Our constructed social knowledge category, along with experiment code and human annotations on some of the SocialIQA data, are released to the research community.[1]

2 Related Work

SocialIQA Task Most previous works on SocialIQA task involve either large size of pre-trained models, and datasets (Khashabi et al., 2020; Lourie et al., 2021) or complicated models that heavily rely on external knowledge bases (Shen et al., 2020; Shwartz et al., 2020; Mitra et al., 2019; Ji et al., 2020a,b; Chang et al., 2020). Among them, UnifiedQA (Khashabi et al., 2020) achieved impressive performance by fine-tuning 11B T5 model (Raffel et al., 2019) with 17 existing QA datasets. Unlike previous efforts, our work achieves comparable performance with a relatively small model and simple knowledge extraction method that does not rely on knowledge bases nor require additional pretraining.

Commonsense Categorization in NLP LoBue and Yates (2011) proposed form-based and content-based categories for commonsense knowledge that is involved in recognizing textual entailment. Boratko et al. (2018) refined the categorization method for knowledge and reasoning proposed for a QA dataset ARC (Clark et al., 2018). The human-annotated relevant sentences was used only for improving the retrieval model, not for the ARC task. In summary none of these work had attempted to leverage such categories in the intended task.

Social Knowledge Categorization Kiesler (1983) proposed a taxonomy for two-dimensional interpersonal behavior, which consists of 16 segments and 128 subclasses. Cowen and Keltner (2017) identified 27 distinct varieties of human emotion, such as anger, excitement, relief, etc. Recently, Forbes et al. (2020) introduced a formalism to study people's social and moral norms over daily life situations, which includes 12 different dimensions of people's judgments. Motivated by these prior work that covered different aspects of knowledge of daily events and social interactions, our work provides a comprehensive overview of social knowledge needed by the SocialIQA task.

3 Methodology

This section presents two approaches to model the underlying semantics and knowledge of the SocialIQA task, together with a simple yet effective method that leverages these knowledge types to improve QA models.

3.1 Question Relation Type

The SocialIQA dataset was derived from the ATOMIC (Sap et al., 2019a). ATOMIC is a knowledge base that focuses on everyday social commonsense knowledge organized as ten types of if-then relations. Based on this observation, we tag each question in SocialIQA according to its relation types in ATOMIC by conducting rule-based mapping between them. Specifically, we match keywords in the questions and use the Spacy model (Honnibal et al., 2020) to detect subjective and objective in context sentences. [2]

Once the mapped types are obtained for each SocialIQA question, we transfer such information to QA models by simply concatenating the tags to original QA examples in the format of $[Context, \text{SEP}, Question, \underline{Tag}, Answer]$ as input to a PLM for fine-tuning. [3] Although we

[1]https://github.com/posuer/social-commonsense-knolwedge

[2]Take the fourth instance in Figure 1 as an example; we firstly match the word “need" in question to the "Need" relation, then detect the name “Taylor" is subjective in the context, so we assign “xNeed" to this question.

[3]All the tags are added into the model's vocabulary as spe-

use RoBERTa model in this work, our method is generic and can be applied to any PLMs.

3.2 Social Knowledge Categorization

Although utilizing ATOMIC to obtain relation type is straightforward and simple, it is restricted to datasets derived from ATOMIC. Inspired by previous work around emotion and social interactions in psychology (Kiesler, 1983; Cowen and Keltner, 2017; Forbes et al., 2020), we propose a taxonomy to categorize social commonsense knowledge into types, which can be generally applied to other social commonsense reasoning related tasks. As shown in Figure 1, this taxonomy includes four categories as follows:

Feelings and Characteristics involves personal feelings and characteristics. Specifically, it includes the development of subsequent feelings and emotions, and events triggered by personal feelings, emotions or characteristics; the feelings and emotions caused by a certain event; and the personal characteristics reflected by a particular event.

Interaction includes events, daily life habits and experiences caused by interactions or emotions among two or more people, as well as interactions and possible responsibilities and obligations between individuals and groups.

Daily Events deal with relationships between daily events, habits, and life experiences. In this category, most situations only involve individuals. Even if multiple people are involved, the focus is on the event itself rather than on the interaction between people. For example, in such a scenario, "Two people went to the hair salon together, what will they do next?", it will be classified into this category even though it involves two persons.

Knowledge, Norm, and Rules Unlike daily events, the events included in this category usually involve social or scientific knowledge and rules that are written in documents or books. The knowledge and rules here may pertain to various topics such as legislation, law, career development, social identity, and medical care.

3.3 SocialIQA-Category Dataset

We manually annotate some of the SocialIQA data with our proposed four social knowledge categories.

cial tokens, for instance $[xNeed]$, and they are concatenated to QA examples in text form.

Model	Dev	Test
RoBERTa large	77.4	77.0
+ **Question Relation**	79.0	78.3
+ **Social Knowledge Cat**egory	79.4	78.5
+ **Question Relation + Social Knowledge Cat**	79.8	78.5
Knowledge Source (Mitra et al., 2019)	79.5	78.0
GLM (Shen et al., 2020)	79.6	78.6
Ablation Study:		
+ Question Relation **(Random)**	77.6	75.3
+ Social Knowledge Category **(Random)**	77.6	76.4

Table 1: Accuracy on SocialIQA test set

Among the total 800 examples in this dataset, 600 training examples are selected from SocialIQA training data, and 200 dev examples are selected from SocialIQA dev data.

Dataset Creation Since the annotation requires fully understanding the social knowledge category, two listed authors annotate two rounds of 50 examples, discuss the disagreements in the middle. The percentage of agreement on the second round is higher than 95%, which indicates that these categories are well-defined. Then each of the two annotators is responsible for another 350 examples separately.

Knowledge Category Prediction We fine-tune RoBERTa large model with the SocialIQA-Category training set and achieve up to 80% accuracy on the dev set for this four-category classification task. The trained model is used to assign category labels to the whole SocialIQA dataset automatically. Follow the same method in section 3.1, the obtained category labels are concatenated to original examples in the format of $[Context, \underline{label}, \text{SEP}, Question, Answer]$ to train and test the QA model.

4 Experiments and Results

We use the SocialIQA data set for our experiments. SocialIQA contains $33,410$ training examples, $1,954$ dev examples and $2,224$ test examples.

Models & Baselines We employ the RoBERTa-large model as baseline. Our proposed method uses RoBERTa-large models in the same way of the baseline, except concatenating the tags or labels to original QA examples (described in Section 3.1 and 3.3). We also compare our methods with the following published models on SocialIQA: *GLM* (Shen et al., 2020) re-trains the RoBERTa

Model	xIntent	xNeed	xAttr	xReact	xWant	xEffect	oReact	oWant	oEffect	Other
RoBERTa-large	0.20	0.22	0.26	0.24	0.26	0.24	0.23	0.23	0.28	0.21
+ Question Relation	**0.19**↓	**0.19**↓	**0.24**↓	**0.23**↓	**0.24**↓	**0.21**↓	0.23	**0.20**↓	**0.26**↓	**0.13**↓
+ Question Relation + Social Knowledge Cat.	0.22	0.23	**0.21**↓	**0.21**↓	**0.21**↓	0.21	0.23	**0.18**↓	**0.25**↓	0.17

Table 2: Error Rate Distribution of *Question Relation* model on SocialIQA Dev Set

Model	Feelings and Characteristics	Interaction	Daily Events	Knowledge, Norm, and Rules
RoBERTa-large	0.25	0.21	0.24	0.21
+Social Knowledge Category	**0.21**↓	**0.19**↓	**0.20**↓	0.22
+ Question Relation + Soc. Knowl. Cat.	0.21	**0.16**↓	0.20	0.26

Table 3: Error Rate Distribution of *Social Knowledge Category* model on SocialIQA-Category Dev Set

model by injecting structured knowledge from the knowledge graph. *Knowledge Source* (Mitra et al., 2019) concatenates the question, answer, and the context as the query to retrieve and re-rank the top ten sentences from ATOMIC, and then fuses them into the QA model to select the right answer.

Training Setup The hyperparameters are selected based on the best performing model on the dev set. We use grid search to fine-tune the model, and select select the learning rate from $\{1e-5, 2e-5\}$, batch size from $\{4, 8\}$ and gradient accumulation from $\{4, 8, 16\}$. The model is trained up to 4 epochs. [4]

Results We report the accuracy results on the SocialIQA test set in Table 1.

- Both ways of using knowledge type information outperform the RoBERTa baseline models. The paired t-test shows that *Social Knowledge Category* achieves significant gains over the RoBERTa model with $p < 0.05$, and *Question Relation* is a little short of significance on test set ($p = 0.06$).
- Compared with *GLM* and *Knowledge Source* that require large amounts of engineering work to explore external knowledge from ATOMIC, our simple and direct utilization of the ATOMIC relations and social knowledge categorization achieves competitive performances.
- The naive combination of question relation type and social knowledge category shows no gains over any single model. One reason may be that these two types of relations are not entirely orthogonal to each other.

[4]Details about the experiment setting is in Appendix.

Error Analysis We conduct detailed error analysis to examine the performance gains from *Question Relation* and *Social Knowledge Category* models. The results are presented in Table 2 and 3.

For the *Question Relation* model, we present the error rate under different relations on the socialIQA dev set. As we can see, the *Question Relation* model achieves consistent improvements on almost all relations, which proves the effectiveness of incorporating the logical relation type.

We also compare the error rate of the *Social Knowledge Category* and the *RoBERTa-large* on the manually annotated dev set with 200 examples. We observe performance gains on all categories except the Knowledge, Norm and Rules category. This suggests that current QA models still struggle with questions involving knowledge and norms, calling for more sophisticated techniques to reason over these social and scientific knowledge. [5]

Ablation Study We conduct the ablation study for our proposed method described in Section 3.1 and 3.3. While training and testing, the mapped relation type tag or predicted knowledge category label is replaced by a randomly chosen tag or label. The experiment result presented in Table 1 shows that the randomly chosen tags or labels could not help the QA models. The performance gains from knowledge of Question Relation or Social Knowledge Category are indeed valid.

5 Conclusion

In this work, using the SocialIQA task as an example, we integrate two different knowledge types into the QA model training: one based on question relations, and the other is our own defined social knowledge category. Experimental results demonstrated that incorporating semantic categorizations of social knowledge into QA models helps boost performances on the social commonsense QA task. The proposed simply ways of incorporating knowledge into the model also achieved comparable performances to these much complicated models.

[5]More examples and error analyses on these two types of relations can be found in Appendix.

6 Ethics

We create a dataset, SocialIQA-Category, by annotating part of the SocialIQA dataset (Sap et al., 2019a). SocialIQA dataset is accessible to the public and can be downloaded from an open URL. All the annotations are done by the listed authors of this paper. The annotations only include the aforementioned relation type and social knowledge category. Our work focuses on QA tasks, specifically the SocialIQA task. Neural models that are created for this task are not supposed to solve any real-world problem. In terms of environmental consequences, all of our experiments are done with the RoBERTa model. Models training for SocialIQA is usually done within 1 hour on a single GPU.

References

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136, Online. Association for Computational Linguistics.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 60–70, Melbourne, Australia. Association for Computational Linguistics.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 74–79.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Alan S. Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

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.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, and Minlie Huang. 2020a. Generating commonsense explanation by extracting bridge concepts from reasoning paths. *arXiv preprint arXiv:2009.11753*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020b. Language generation with multi-hop reasoning on commonsense knowledge graph. *arXiv preprint arXiv:2009.11692*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Donald J Kiesler. 1983. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological review*, 90(3):185.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 329–334.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *arXiv preprint arXiv:2103.13009*.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2925–2933.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. How additional knowledge can improve natural language commonsense question answering? *arXiv preprint arXiv:1909.08855*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting structured knowledge in text via graph-guided representation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8980–8994.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*.

A Appendix: Experiment Setting

The number of hyperparameter search trials is 12 for each model. During training, each epoch takes about 30 mins on average. The hyperparameter configurations for best-performing models are listed as below.

Model	Learning Rate	Batch Size	Gradient Accumulation
RoBERTa Large	$1e-5$	8	8
+Question Relation	$1e-5$	8	8
+Social Knowkedge Cat	$1e-5$	8	4
+Question Relation + Social Knowledge Cat	$1e-5$	8	16

B Appendix: Case Study

We conduct a case study to understand how the two types of categorical information help the QA task. The examples are listed in Table 4, 5.

Question Relation Type: **xNeed**
Riley told Austin's landlord that Austin was making a lot of noise at late hours.
What does Riley need to do before this?
(a) potentially call the police (b) document incidents ✓ | **Question Relation** (c) follow up with the landlord | **RoBERTa**

Table 4: Case Study for Question Relation (✓: golden label)

Social Knowledge Category: **Feelings and Characteristics**
Austin was taking a test and found it difficult at first.
How would you describe Austin?
(a) student | **RoBERTa** (b) stupid (c) overwhelmed ✓ | **Social Knowledge Category**

Table 5: Case Study for Social Knowledge Category (✓: golden label)

B.1 Question Relation Type

Table 4 refers to an example that is predicted correctly by the *Question Relation* model but wrongly by the RoBERTa model. The RoBERTa model mistakenly selects the answer describing an event that could happen after the noise complaint. However, the question is asking possible events before it. The question type "xNeed" exactly provides such signal and indicates that the right answer should happen before the given context, and thus helps the QA model choose the right answer.

B.2 Social Knowledge Category

In Table 5, although the RoBERTa model selects the reasonable answer "student" to describe Austin, it fails to infer more in-depth semantic information embedded in the context. The actual focus is that Austin found the test challenging and was overwhelmed. Our social knowledge taxonomy assigns this example to the "Feelings and Characteristics" category. It enables the QA model to pay more attention to candidates that emphasize a kind of personal feeling.