Investigating the Generative Approach for Question Answering in E-Commerce

Kalyani Roy[1], **Vineeth Kumar Balapanuru**[1], **Tapas Nayak**[2*] and **Pawan Goyal**[1]
[1]Indian Institute of Technology Kharagpur, India
[2]TCS Research, India
kroy@iitkgp.ac.in, Vineethkumar6001@gmail.com
tnk02.05@gmail.com, pawang@cse.iitkgp.ac.in

Abstract

Many e-commerce websites provide Product-related Question Answering (PQA) platform where potential customers can ask questions related to a product, and other consumers can post an answer to that question based on their experience. Recently, there has been a growing interest in providing automated responses to product questions. In this paper, we investigate the suitability of the generative approach for PQA. We use state-of-the-art generative models proposed by Deng et al. (2020) and Lu et al. (2020) for this purpose. On closer examination, we find several drawbacks in this approach: (1) input reviews are not always utilized significantly for answer generation, (2) the performance of the models is abysmal while answering the numerical questions, (3) many of the generated answers contain phrases like "I do not know" which are taken from the reference answer in training data, and these answers do not convey any information to the customer. Although these approaches achieve a high ROUGE score, it does not reflect upon these shortcomings of the generated answers. We hope that our analysis will lead to more rigorous PQA approaches, and future research will focus on addressing these shortcomings in PQA.

1 Introduction

With the increase in e-commerce shopping, customer-generated product queries are also growing. Manually answering the questions in real-time is infeasible, and also some questions go unanswered for an extended period. It is necessary to answer the user queries in the e-commerce business automatically. The user reviews are a vast source of information with diverse opinions, and they can be used to answer user queries. Earlier works on product question answering (PQA) focus on retrieval-based approaches and binary answer prediction tasks. McAuley and Yang (2016); Fan et al. (2019); Yu and Lam (2018) aim to predict the answer as "yes/no" based on the relevant reviews, customer ratings, aspects in the reviews, etc. Retrieval-based approaches try to find the most relevant review snippet as the answer (Chen et al., 2019a) and use a ranked list of review snippets as the response for a given question (Yu et al., 2018). With the success of machine translation (Sutskever et al., 2014) and summarization (See et al., 2017), the PQA approaches are shifting towards natural answer generation from relevant product reviews (Gao et al., 2019; Chen et al., 2019b; Deng et al., 2020; Lu et al., 2020; Gao et al., 2021). In this work, we analyse the answer generated from state-of-the-art generative models OAAG(Deng et al., 2020) and CHIME(Lu et al., 2020) in detail beyond their traditional scores on popular metrics such as ROUGE (Lin, 2004). We find that despite achieving a good score on these metrics, generated answers have several drawbacks that can lead to user dissatisfaction.

2 A State-of-the-art Generative PQA Model

2.1 PQA Dataset

The Amazon Question Answering dataset (McAuley and Yang, 2016) contains around 1.4 million questions from different categories with multiple customer-written answers and opinion labels, such as positive, negative, and neutral. The Amazon Product Review dataset (Ni et al., 2019; He and McAuley, 2016) includes users' reviews along with a rating of the product given by the same user. The Product ID is used to align the question with its reviews.

2.2 PQA Models

We use Opinion-aware Answer Generation (OAAG) model (Deng et al., 2020) and Cross-passage Hierarchical Memory Network (CHIME)

[*] This work was carried out while author was a postdoctoral researcher at IIT Kharagpur.

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 210 - 216

model (Lu et al., 2020) for our analysis. Following the generative approach, these two models achieve state-of-the-art performance on the Amazon Question Answering dataset. There are thousands of products in each category in the Amazon Product Review dataset, and each product has thousands of reviews. All the reviews may not be relevant for a particular query, and therefore, to answer a product-related question, models need to filter out the irrelevant reviews first. OAAG and CHIME use the BM25 algorithm to retrieve and rank all the review snippets of a product, and the top k relevant snippets (we use top 10 reviews snippets) for that question are taken as the premise of the answer.

2.2.1 OAAG Model

Upon retrieving the relevant reviews, OAAG uses an encoder-decoder model for answer generation. OAAG encodes the question and each review corresponding to that question using a Bi-LSTM (Hochreiter and Schmidhuber, 1997) network. They apply a co-attention mechanism over these encodings to get the question and review representations. They utilize the ratings of the retrieved reviews to mine the general opinion about the question using the attention mechanism. Finally, they employ a multi-view pointer-generator network that copies words from the question as well as from the reviews and fuses the opinion by re-weighting the attention scores of the words in reviews to generate an opinionated answer. They report ROUGE-based scores to compare the model performance against the previous approaches (Chen et al., 2019b; Gao et al., 2019).

2.2.2 CHIME Model

CHIME uses a transformer-based encoder-decoder model to generate the response. It extends pretrained XLNet (Yang et al., 2019) with an auxiliary memory module that consists of two components: the *context memory*, and the *answer memory*. Given a question with K review passages, it creates K training instances, each consisting of the question, a review passage, and the reference answer. Each training instance is fed into an XLNet encoder to get the hidden representations that are used to update the two memories. The *context memory* mechanism sequentially reads the review passages and gathers the cross-passage evidences to identify the most prominent opinion in reviews. The *answer memory* works as a buffer to gradually refine the generated answers after reading each (*question, review passage*) pair. After reading the last review, the *answer memory* is fed to the decoder to get a final response.

3 Research Questions

We empirically analyse the OAAG model with dynamic fusion and CHIME model to answer the following research questions:
RQ1 : Are the retrieved review snippets significantly utilized for generating the answers?
RQ2 : Is the model performing similarly for a heterogeneous group of questions?
RQ3 : Is the generative model biased towards more frequently occurring phrases?
RQ4 : Can ROUGE capture the correctness of generated answers?

4 Experiments

We use two product categories, namely, *Home&Kitchen* and *Sports&Outdoors* for our analysis from the dataset mentioned in Section 2.1 after combining the question-answer and review dataset with the Product IDs. We will denote the two categories as *Home* and *Sports*, respectively. We use the same data split from OAAG[1] to retrain the models. Since there is no validation dataset, we take the 10% of the train data as validation data. Table A.1 in the Appendix shows the details of training, validation, and test split. We keep all the hyper-parameters the same as the OAAG and CHIME. We train all the OAAG models for 20 epochs and CHIME models for 3 epochs, and the model that performs the best on the validation set is used to evaluate the test set. We evaluate the model with ROUGE metric and report the F1 scores for ROUGE-1 (R1) and ROUGE-L (RL), which measure the word overlap and the longest common sequence between the reference answer and the generated answer, respectively. We obtain the ROUGE scores using `rouge-score` [2] package.

5 Analysis & Discussion

5.1 Answer to RQ1 (Utilization of retrieved review for generating the answers)

Both the models use the BM25 algorithm to retrieve relevant reviews using the questions in the test dataset. We refer to this test setting as *BM25Q*.

[1] https://github.com/dengyang17/OAAG
[2] https://pypi.org/project/rouge-score/

For answering RQ1, at inference time, we replace these reviews with four sets of review snippets: (i) *TrainA*: We use BM25 to find the closest question to the test question in the train data, and we take the answer of it as the generated answer. (ii) *RandomOD*: We randomly choose the review snippets from any other product of that category except the product for which the question is asked. (iii) *RandomID*: We randomly select review snippets from the review sentences of that particular product. (iv) *BM25QA*: We retrieve the review snippets using the BM25 algorithm that uses the question and reference answer in the test dataset.

OAAG uses the opinion along with the reviews. We also select the opinion of the corresponding review sentence while replacing the reviews. Both the models utilize the top 10 reviews for training and evaluation.

		Sports		*Home*	
		R1	RL	R1	RL
	TrainA	13.01	10.13	14.36	11.35
OAAG	BM25Q	15.01	**11.99**	14.44	11.91
	RandomOD	14.25	11.38	14.04	11.53
	RandomID	14.71	11.69	14.42	11.85
	BM25QA	**15.09**	11.97	**14.53**	**11.93**
CHIME	BM25Q	**18.53**	**13.19**	**18.99**	**13.84**
	RandomOD	18.10	12.87	17.83	13.11
	RandomID	17.95	12.81	17.98	13.17
	BM25QA	17.99	12.84	17.85	13.11

Table 1: Performance of the OAAG and CHIME models with various sets of review snippets.

Table 1 shows the result of this experiment. The TrainA does not utilize either of the models to generate the answer. It shows the answer from the most similar train question, and its performance is competitive with other methods, especially in *Home*. In both the categories, the performance of both the models is almost similar in RandomOD and RandomID. RandomID shows marginally better performance than RandomOD for OAAG. For CHIME, BM25Q performs the best in both categories. For OAAG, BM25QA performs the best in *Home* while in *Sports*, BM25QA performs the best in R1, and BM25Q performs the best in RL, but the difference is minute. The results are quite surprising: the performance of the models is very similar when the answers are generated with random reviews vs. when the answers are generated with the reviews obtained from BM25. Hence, it is not clear if the model is effectively utilizing the retrieved review snippets.

5.2 Answer to RQ2 (Models' performance on heterogeneous questions)

Different types of questions are asked on the Amazon product page like numerical, "yes/no", descriptive. The generative model may not be suitable for answering all kinds of questions. So, we categorize the questions as template-based and descriptive.

		Sports		*Home*	
		R1	RL	R1	RL
OAAG	Template	13.15	10.99	12.38	10.33
	Descriptive	**15.67**	**12.34**	**15.11**	**12.21**
CHIME	Template	16.72	12.79	17.68	13.67
	Descriptive	**19.17**	**13.33**	**19.37**	**13.89**

Table 2: Performance of OAAG and CHIME models on template-based, descriptive categories of questions.

For template-based questions, the answer can be yes or no without any explanation. We filter the questions where the answer starts with 'yes', 'yeah', 'no', 'nope' and mark these as template-based questions. Both categories contain $\sim$75% descriptive questions. Table 2 summarizes the result of the template-based and generative questions. Both models' performance in descriptive questions is better than the template-based questions.

Furthermore, we categorized the questions into numerical and non-numerical questions. We consider a question to be numerical if there are numbers in the question or in the reference answer. The test datasets of both the categories have $\sim$19% numerical questions. The OAAG model performs better in answering non-numerical questions, while CHIME performs better in answering numerical questions. Although the ROUGE scores are close in numerical and non-numerical questions for both the models, on analyzing the numerical answers, we find that the words in generated and reference answers might match, but the numbers generally do not match. [3] We present some examples of numerical questions with their answers in Table A.4 of Appendix.

5.3 Answer to RQ3 (Bias in model)

We observe that some phrases are frequently occurring in the reference answers as well as in the generated answers. We find that in the training data of both categories, $\sim$2.4% of the reference answers

[3] We manually check 400 numerical question answers for OAAG, and only 2 answers turn out to be correct. We check 100 random numerical question answers for CHIME, but none are correct.

start with the phrase "I don't think so", but 12.29% of responses in *Sports* and 35.64% responses in *Home* begin with this phrase. This ∼2.4% repetition of the same phrase in the training data makes the generative model biased towards this phrase.

		Sports		*Home*	
		R1	RL	R1	RL
OAAG	BM25Q	**15.01**	**11.99**	14.44	11.91
	BM25Q-IDK	14.87	11.71	**16.16**	**12.73**
CHIME	BM25Q	**18.53**	**13.19**	18.99	13.84
	BM25Q-IDK	18.50	13.21	**19.44**	**14.12**

Table 3: Performance of OAAG and CHIME models when trained with and without phrases that are not meaningful (IDK phrases).

Many of the reference answers in the training data contain "I don't know", "I have no idea", "I can't say". These kinds of answers do not give any meaningful information to the user. Together, we denote these phrases as IDK. On analysis of the dataset, we find that in *Sports*, there are 3.04%, 2.9%, and 6.9% IDK phrases in train dataset answers, test dataset answers, and generated answers, respectively. In *Home*, the answers in the train and test dataset contain 3.64% and 3.60% IDK phrases, respectively, but 16.31% of the answers are generated as IDK phrases. So, in the generated answers, the appearance of IDK phrases has increased by three to five folds which clearly shows that the model is biased towards these frequent phrases.

To see the effect of these phrases on the models, we remove the questions from the training dataset which have IDK in their reference answer [4] and retrain the models. We denote this model as BM25Q-IDK. Table 3 shows the result of BM25Q, the model trained on the original training data, and BM25Q-IDK. *Home* had 16.31% and *Sports* had only 6.9% IDK phrases, and consequently, when the IDK phrases are removed, it has more impact on *Home* which drops the bias towards these phrases and improves the ROUGE score, whereas, in *Sports*, BM25Q and BM25Q-IDK have close ROUGE scores due to lesser IDK phrases in the generated answer.

5.4 Answer to RQ4 (Correctness of generated response)

For answering RQ4, we look into the generated response with high R1 scores and check their correctness with respect to the reference answer. In OAAG model, 15.36% predictions in *Home* and 13.34% predictions in *Sports* have R1 score above 20. We manually analyse the reference and generated answers of randomly chosen 100 question-answers with a high ROUGE score (> 20), and we find that 54% are answered incorrectly. In CHIME model, 46.87% predictions in *Home* and 46.15% predictions in *Sports* have R1 score above 20 and 56% of 100 randomly chosen question-answer pairs (whose ROUGE score > 20) turn out to be incorrect.

Question	is this box made of polypropylene? can photos and mementos be safely stored in it ?
Ref. Ans.	i can't comment on the material it is made of, but the top does not stay on tight. it is not a decent storage container for anything requiring a top or to be airtight.
OAAG.	it says "made in china" on the bottom of the box . it says "made in china" on the bottom of the box . hope that helps . (26.22)
CHIME.	yes, it is polypropylene and can be stored in the box if you want to store them in a safe environment that is not toxic to you and/or your pets or food they will be exposed (27.77)
Question	What is the width at the base?
Ref. Ans.	Width across the bottom/base is approximately 3 inches.
OAAG.	The width of the top of the base is about 1.5 inches . the base of the pitcher is 9 inches. the top of the pitcher is about 1.5 inche.' (25.00)
CHIME.	itś about 12 " wide at the base and about 10.5 inches deep (26.08)

Table 4: Examples of generated answers with high R1 score, but having incorrect answer. The R1 score is mentioned within the parenthesis.

Table 4 shows two examples where the generated responses result in high R1 scores, but the answers are incorrect. Both models predict irrelevant answers in the first question, and the predicted dimension is incorrect in the second question. It shows that it is not possible to infer from ROUGE scores if the generated answer is accurate to the reference answer, i.e., the word count overlap is not an indicator of an accurate answer. We show some more cases with high R1 scores in Tables A.2 and A.3 in the Appendix.

6 Conclusion

In this paper, we extensively analyze the generative approach of question-answering in e-commerce using a state-of-the-art OAAG model (Deng et al., 2020) and CHIME model (Lu et al., 2020). We find many shortcomings which need to be addressed for a reliable PQA system. We try to address four re-

[4] Phrases such as "I don't think so" can be considered as valid reference answers. Therefore, we do not remove it from the training data.

search questions related to the generative approach for PQA, such as how the models utilize the reviews, how it performs on diverse question types, whether it is biased toward frequent phrases in training data, and the correctness of the generated response. We hope that our analysis will lead to more rigorous PQA research.

References

Long Chen, Ziyu Guan, Wei Zhao, Wanqing Zhao, Xiaopeng Wang, Zhou Zhao, and Huan Sun. 2019a. Answer identification from product reviews for user questions by multi-task attentive networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019b. Review-driven answer generation for product-related questions in e-commerce. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*.

Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Opinion-aware answer generation for review-driven question answering in e-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Miao Fan, Chao Feng, Mingming Sun, Ping Li, and Haifeng Wang. 2019. Reading customer reviews to answer product-related questions. In *Proceedings of the 2019 SIAM International Conference on Data Mining*.

Shen Gao, Xiuying Chen, Z. Ren, Dongyan Zhao, and Rui Yan. 2021. Meaningful answer generation of e-commerce question-answering. *ACM Transactions on Information Systems*.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics.

Junru Lu, Gabriele Pergola, Lin Gui, Binyang Li, and Yulan He. 2020. CHIME: Cross-passage hierarchical memory network for generative review question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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)*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Qian Yu and Wai Lam. 2018. Review-aware answer prediction for product-related questions incorporating aspects. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

Qian Yu, Wai Lam, and Zihao Wang. 2018. Responding E-commerce product questions via exploiting QA collections and reviews. In *Proceedings of the 27th International Conference on Computational Linguistics*.

A Appendix

A.1 Dataset Statistics

We downloaded the dataset from the Github repository of OAAG (https://github.com/dengyang17/OAAG). We use the same data split, but as there were no validation dataset, we use 10% train data as validation data. The deails of training, validation, and test split is shown in Table A.1.

	Sports	*Home*
Train	40,517	73,118
Validation	4,502	8,125
Test	5,001	9,026

Table A.1: Dataset statistics.

Question:	does the sight come with batteries ? if not what batteries do i need to buy ?
Ref. Ans.:	the sight did come with a battery , but if needed it would be a cr2032 .
Gen. Ans.:	no , it does not come with a batteries . you need to buy a separate batteries . it 's a great scope for the money but it 's not worth it . (31.82)
Question:	is it lined ?
Ref. Ans.:	it does n't have a liner in it , but still makes for a very nice light winter coat ...
Gen. Ans.:	no , it is not lined . it is a very nice jacket and i do n't think it 's too big for it . it 's a great deal . (22.22)
Question:	does the bottle sweat ? and , does the straw have a cover ?
Ref. Ans.:	does it sweat ? a little . not hardcore . just a little condensation on the bottom . and the straw does n't have a cover ...
Gen. Ans.:	no , it does not have a sweat straw . it is a very good straw and i have n't had any issues with it . i have n't had any issues with it . (24.62)
Question:	is it wood ?
Ref. Ans.:	it is gorgeous , and yes it is wood
Gen. Ans.:	no it is not wood . it 's a very nice piece of furniture . i would n't buy it again . it 's a very nice piece of furniture . (22.22)
Question:	how many watt ?
Ref. Ans.:	i do n't know but i am disappointed in mine . i wanted it to keep my coffee hot but it only keeps it warm .
Gen. Ans.:	i do n't know , but i do n't know if it does n't work for me . i 've had mine for over a year and it works great . (39.29)

Table A.2: Examples of answers generated by the OAAG model with high R1 scores. The R1 scores are within the parenthesis.

A.2 Generated Answers with High R1 Score

Table A.2 and Table A.3 shows some examples where the generated answers have a high (> 20) R1 score. In Table A.2, in the first and the second example, the generated answers are exactly the opposite of the reference answers. In the third example, the question was about sweating of the bottle and straw cover, but the answer does not address any of these. In the fourth example, the answer is ambivalent. The last example contains a frequently occurring phrase "I don't know" with a very high R1 score.

Similarly, in the case of Table A.3, the second and the fifth examples have high R1 scores, but the generated answers are exactly opposite of the reference answers. In the first question, the model wrongly predicts that it would melt and bubble up a little in the microwave, and in the third question, it predicts an entirely different answer. The response to the fourth question is "I don't know," which is a frequently occurring phrase.

Question:	will it melt or bubble in the microwave ?
Ref. Ans.:	it never has with me and i have a high power microwave. really love to freeze in them. no ice buildup in them .
Gen. Ans.:	i have not tried it in the microwave but it does melt and bubble a little when you press on. (32.55)
Question:	does this actually comes with storage box as shown which is not in description ?
Ref. Ans.:	mine did not come with a storage box. and, it was not a very good knife. it grinds rather than cuts.
Gen. Ans.:	yes, it comes with a storage box as shown in picture.it is very easy to use and easy to take apart and put back together (38.29)
Question:	does it have an attachment to clean windows?
Ref. Ans.:	there is a wider flat piece that you could use on windows... it works nicely .
Gen. Ans.:	no, it does not have an attachment to clean windows but you can use a hose to do so if you need to clean them with water or something similar like that in your kitchen .(20.40)
Question:	could i have a red one ?
Ref. Ans.:	they do come in red, white and cream colors. whoever you buy it from, email them right after you place your order so you have a better chance of getting red .
Gen. Ans.:	i don't know if you can get a red one but it is a great product and is easy to use with any size pasta or sauce you want to make in your kitchen . (27.27)
Question:	is the under side a refrigerator as well or just storage ?
Ref. Ans.:	its a storage not a refrigerator you can only store cups or water bottles .
Gen. Ans.:	it's a refrigerator, but you can put it in the freezer or in your car to cool off water and then use it for storage .(35.00)

Table A.3: Examples of answers generated by the CHIME model with high R1 scores. The R1 scores are within the parenthesis.

Question:	why do the dimensions say $2''$ when the description says $3''$? is the thickness of the mattress 3 inches or 2 ?
Ref. Ans.:	$3''$ .. seems amazon just copy and pasted from the other products in the line-up .
OAAG:	the thickness of the mattress is $3''$ and the $3''$ $3''$ is the same as the picture of the mattress . it is a very nice pillow .
CHIME:	it's 3 " and 2.5 inches thick.i love this bed and it is perfect for my son who sleeps on his bed with no problem
Question:	what is the width and length of the rubbermaid 5576 slide and stack ?
Ref. Ans.:	the outside measurements are : 14 in . wide by 20 in . long . the inside of the basket is $123/4$ in . wide by $163/4$ to $181/2$ in . long ...
OAAG:	it 's about $14''$ long by $12''$ wide by $12''$ deep by $4''$ high . hope that helps . i am very happy with this purchase .
CHIME:	the dimensions of the slide and stack are 11.5 inches wide by 8.5 inches deep by 7.75 inch tall.
Question:	how big of a pot can you put on it , $12''$?
Ref. Ans.:	the cooking top measures $10.5'' \times 10.5''$. the burner is an $8''$ diameter circle . a $12''$ pot would extend over the control panel if it were centered .
OAAG:	i have n't measured it , but i do n't think it would be too big to fit in a pot . i use it for that purpose and it works great .
CHIME:	it's 12 inches long and 12 inches wide and about 2 inches deep at the bottom of the pans.
Question:	what is the exact weight of it and how strong is the pocket clip ?
Ref. Ans.:	it 's about 3 oz . and it 's just as strong as any other clip i 've seen
OAAG:	it 's about $7.5''$ long . it 's a great knife . it 's a great deal . it 's a great knife and i love it . it 's a great deal .
CHIME:	i'm not sure about the exact weight but it is very light and is very easy to use with ease .
Question:	does this fit a 2 year old ?
Ref. Ans.:	it fit my two year old who has a pretty normal size head . the multiple thicknesses of pads that are included really make it a great long term helmet !
OAAG:	yes it will , it 's a very tight fit . i do n't think it would be too big for a $2''$ $2''$. it 's a great helmet .
CHIME:	i bought this for my son and he loves it so much he bought another one for his 2 year-bean .

Table A.4: Examples of numerical questions with their reference answer and the answer generated by OAAG and CHIME models.

A.3 Examples of Numerical Questions

We present some examples of numerical questions with their answers in Table A.4. In the first example, the generated answer is right, but none of the answers are correct for the rest of the questions.