Instant Answering in E-Commerce Buyer-Seller Messaging using Message-to-Question Reformulation

Besnik Fetahu[†], Tejas Mehta[†], Qun Song, Nikhita Vedula, Oleg Rokhlenko, and Shervin Malmasi

Amazon.com, Inc. Seattle, WA, USA
{besnikf,mehtejas,qunsong,veduln,olegro,malmasi}@amazon.com

Abstract. E-commerce customers frequently seek detailed product information for purchase decisions, commonly contacting sellers directly with extended queries. This manual response requirement imposes additional costs and disrupts buyer's shopping experience with response time fluctuations ranging from hours to days. We seek to automate buyer inquiries to sellers in a leading e-commerce store using a domain-specific federated Question Answering (QA) system. The main challenge is adapting current QA systems, designed for single questions, to address detailed customer queries. We address this with a low-latency, sequence-to-sequence approach, MESSAGE-TO-QUESTION (M2Q). It reformulates buyer messages into *succinct* questions by identifying and extracting the most salient information from a message. Evaluation against baselines shows that M2Q yields relative increases of 757% in question understanding, and 1,746% in answering rate from the federated QA system. Live deployment shows that automatic answering saves sellers from manually responding to millions of messages per year, and also accelerates customer purchase decisions by eliminating the need for buyers to wait for a reply.

1 Introduction

The rapid growth of e-commerce, with hundreds of millions of global users, hinges on customer satisfaction tied to easy, instant access to comprehensive product details. However, the available product descriptions, customer reviews, and QAs on e-commerce sites may not provide enough information for users to make informed purchasing decisions.

E-commerce platforms use a *Seller Messaging* feature for facilitating communication between buyers and sellers through an online messaging system about product details, shipping, or post-purchase concerns. However, large platforms which handle over 100K messages daily experience high wait times ranging from *1-2 hours* to *several days* due to the time and monetary costs of manual responses, thus delaying purchase decisions.

We address these gaps by leveraging a federated, multiple backend QA system for instantly answering customer questions using knowledge sources such as reviews, product catalogs, manuals, and community QA. However, *style mismatch* is a key challenge: customers write lengthy email-style messages, while traditional QA systems operate on simple direct questions. We tackle this via an end-to-end approach, MESSAGE-TO-QUESTION (M2Q), using sequence-to-sequence models to reformulate

[†] These authors contributed equally to this work.

Buyer's Message to Seller	Ground Truth	Reformulated Message
I'm trying to look into Forids trash bags, and have found that the website and facebook on the box dont exist :(***I'm curious about their sourcing and material is used*** !!	what is the sourcing material of this product?	*what is the sourcing material used for this product*?
Hello! I'm curious whether ***this Re:Zero REM figure you're selling is from the authentic Taito Coreful brand?*** Thank you.	is this product from the taito coreful brand?	*is this product from the authentic taito coreful brand?*
My family surname is not a known surname. Are you ***able to create a family crest with all the emblems and mottos?*** Looking forward to hear from you.	can this product be customized with all the emblems and mottos?	*can this product be created with all the emblems and mottos*?

Table 1: Examples of buyer messages reformulated by M2Q with the buyer's main intent in red.

lengthy buyer messages into short standalone questions. M2Q is optimized to distill relevant details from buyer's messages into *answerable* questions and use a state-of-the-art QA model to generate precise answers. This allows the Seller Messaging feature to provide instant QA by leveraging existing resources without incurring additional time or cost overheads (see Table 1). During message reformulation, M2Q considers the buyer's primary needs, combines multiple needs if necessary into a concise request for the QA system, and ensures customer privacy by omitting personal information.

2 Background

Buyer-Seller Interactions & QA: Prior research has highlighted the financial and commercial importance of studying interactions between buyers and sellers on e-commerce stores, as well as answering buyer inquiries in a fast and effective manner [18,13,1]. We propose to improve buyers' access to instantaneous answers, and reduce product sellers' burden and expenses on e-commerce stores by integrating product question answering (QA) systems [12,7] into the Seller Messaging feature. Prior work on automatically answering customer queries within customer service applications, focuses on retrieving answers from a knowledge base [14,6,15,23] as well as generating answers [3,20].

Text Reformulation: Buyer messages can be verbose with long descriptions or irrelevant personal details (see Table 1). This distracts QA models and makes it difficult for them to provide accurate answers [16,2,24]. M2Q ensures that buyer message reformulations can *adapt* to and maximize the understanding and answering rate of existing QA systems. Several Large Language Models (LLMs) have been fine-tuned to summarize and extract salient information from dialogues and email threads [10,22]; identify the appropriate context to be input to a model to answer questions effectively [19,17]; reformulate questions for easy answering [11,9,25]; as well as to select the most relevant conversation history as context in case of conversational QA [27,8].

3 M2Q: Message-to-Question

M2Q reformulate buyer's messages into succinct questions that are instantly answered using a federated QA system. If the question cannot be answered, or buyers are dissat-

isfied with the automatic response, they can forward their message to the seller for a manual response. Figure 1 shows an overview of the approach. Messages sent to sellers are reformulated into questions, which is then used to retrieve an instant answer from the QA system. Next, we describe the components from the figure.

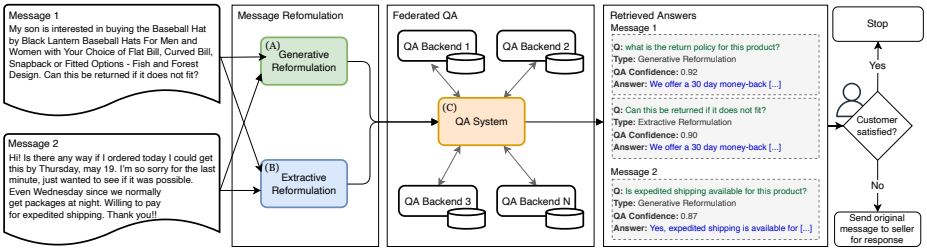

Fig. 1: M2Q approach overview. For each buyer message that may contain multiple questions/intents, M2Q reformulates them (using (B) and (C)) into a *succinct question* with the most salient intent (or conjunction of intents). These are sent to a federated QA system (C) for instant answering. If customers are unsatisfied with the response or receive none, the original message is forwarded to the seller for manual reply.

Generative Reformulation (A): The main objective of this step is to convert buyers' lengthy messages into a shorter, concise form that correctly captures their primary intent, filtering out any irrelevant details or personal information. Table 1 shows examples of such reformulations. For the generative reformulation, we fine-tune different sequence-to-sequence models to perform the rewriting automatically, using a parallel dataset of buyer messages and their human generated reformulations (c.f. §4).

Extractive Reformulation (B): A simpler reformulation method involves splitting the buyer message into sentences, and running them through a question classifier provided by the QA system. The sentence with the highest confidence is chosen to represent the entire message. This method works well for simple messages containing a direct question, but fails on more complex messages, and may not identify any questions at all.

Question Answering System (C): The reformulated messages from the previous steps are input to a federated QA system, which passes them to numerous in-house answer retrieval systems. This system returns answers and a *QA confidence score*. The QA system is a black box in this work; we do not modify or tune it. We show in §5.3 that shorter messages reformulated by M2Q maximize the performance of downstream QA systems, leading to an increase in the automatic question answering rates.

4 Dataset Construction

We created a dataset of ∼6k pairs of buyer messages and their target reformulations, divided into 5k/600/450 for train/dev/test sets. Most messages are between 25-75 words.

Annotation Guidelines. We engaged in-house annotators to transform messages into answerable questions based on certain guidelines, ensuring the quality of the parallel dataset[†]. First, they checked for English language and then identified the number of intents in each message. With multi-intent messages, they focused on the primary intent considering the context, while the remaining were most likely follow-up questions. The message was transformed into a precise question that pinpointed the buyer's main intent and information need. Table 1 shows example annotations.

Annotation Quality Control. A secondary quality check is performed on all annotations by an expert annotator due to the subjective nature of summarizing buyer messages and potential for misinterpretation. The absolute agreement rate between expert and in-house annotators is 76%. Discrepancies are corrected by the expert annotator.

5 Offline Experiments and Results

5.1 Baselines and Our Approach

We compare several reformulation methods[†] against the same QA system.

Extractive Reformulation: The sentence with the highest question classifier confidence score from the buyer message is selected. This works best with short messages that contain a clear, single question.

Flan-T5-XXL: We assess FLAN-T5's reformulations in zero- and few-shot (3 exemplars) settings.

Vicuna-13B: We assess the Vicuna LLM [4], which reportedly obtains 90% of GPT-4's performance, in the same zero-shot and few-shot settings as FLAN-T5.

Approach: We evaluate our M2Q approach with two underlying base models. M2Q-T5 uses the smaller 220M parameter `T5-base` model [21], while M2Q-FT5 uses FLAN-T5-XXL [5] with 11B parameters. We consider two configurations:

- **M2Q:** The buyer's original message is converted into a standalone question via generative reformulation only, and sent to the QA system.
- **M2Q-HYBRID**: We combine generative and extractive reformulation methods. The extractive approach responds when it obtains an answer, else generative reformulation is employed. M2Q-HYBRID thus allows direct responses to shorter messages, decreasing error probability from reformulations.

[†] Our human annotators are expert in-house annotators that provide their relevance judgements based on a pre-determined annotation protocol, which was designed specifically for this task.

[†] Due to privacy regulations, we cannot use external API-based LLMs like ChatGPT.

5.2 Evaluation Strategies

We measure reformulation quality in several ways:

- **Text Generation Performance**: We use BLEU and ROUGE to measure how closely the revised messages align with their originals.
- **Reformulation Accuracy:** Human annotators evaluate reformulated messages based on binary score relevance, verifying the buyer's true intent.
- **Understand Rate:** Measures the QA system's capability to understand buyer messages or the corresponding reformulations.
- **Answer Rate:** Measures the answerability of buyer messages by our QA system.
- **Answer Relevance:** We manually rate each message or its revision as either "*helpful*" or "*unhelpful*" based on the answer's correctness or if it's unanswered.

5.3 Results

		Generation Performance							QA Performance		
	Reformulation Accuracy	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE1	ROUGE2	ROUGEL	Understand Rate	Answer Rate	Answer Relevance
Extractive Baseline	41%	0.143	0.060	0.036	0.025	0.228	0.099	0.214	-	-	-
VICUNA-ZERO-SHOT	-	0.139	0.058	0.037	0.027	0.330	0.121	0.292	+235%	+827%	+320%
VICUNA-FEW-SHOT	31%	0.335	0.206	0.145	0.098	0.504	0.280	0.484	+578%	+1,182%	+1,280%
FT5-ZERO-SHOT	-	0.356	0.182	0.136	0.108	0.489	0.231	0.450	+616%	+1,246%	+1,240%
FT5-FEW-SHOT	58%	0.390	0.210	0.153	0.117	0.520	0.265	0.489	+651%	+1,282%	+1,300%
M2Q-T5	79%	**0.547**	0.369	0.295	0.243	**0.606**	0.384	**0.586**	+746%	+1,478%	+1,820%
M2Q-FLAN-T5	82%	0.546	**0.394**	**0.319**	**0.273**	0.599	**0.406**	**0.586**	+755%	+1,727%	**+2,220%**
M2Q-HYBRID-T5	-	-	-	-	-	-	-	-	+749%	+1,500%	+1,860%
M2Q-HYBRID-FLAN-T5	-	-	-	-	-	-	-	-	**+757%**	**+1,746%**	**+2,220%**

Table 2: We report reformulation performance (left side of the table) measured in terms of reformulation accuracy, BLEU and ROUGE, and QA performance.

Generation Performance: Table 2 shows the BLEU and ROUGE metric results. Both FT5 and VICUNA, achieve lower performance than M2Q for both zero-shot and few-shot scenarios. For FT5 in the few shot setting the exemplars help to obtain better generation performance with an improvement of 3.4 BLEU-1 points. While VICUNA underperforms FT5, few-shot exemplars allow it to significantly improve performance, with an increase of 19 BLEU-1 points. Both LLMs are significantly outperformed by M2Q, with an increase of 9 ROUGE-L points.

The M2Q results highlight two factors: (1) generative models are better for this task, as the extractive baseline achieves the lowest performance on all metrics. (2) reformulating buyer messages is complex, and accurate performance requires fine-tuning.

Reformulation Accuracy: Due to limitations of automated metrics [26], we manually assess reformulation accuracy, computed for: VICUNA–FS, FT5–FS, M2Q-T5 and M2Q-FT5. M2Q obtains the highest reformulation accuracy with 79% for M2Q-T5 and 82% for M2Q-FT5. Note that the size of the base model is important: M2Q-FT5 has 3% higher accuracy than M2Q-T5. Without fine-tuning, FT5 only obtains a 58% accuracy, a drop of ▼24% compared to M2Q-FT5. Similarly, VICUNA obtains a reformulation accuracy of 31%, a ▼51% drop compared to M2Q-FT5.

Question Answering Performance: Table 2 shows the QA results in terms of the understanding confidence scores and answer rates.† For reasons of confidentiality, the QA results are reported as relative improvements over the extractive baseline. On question understanding, M2Q-T5 and M2Q-FT5 obtain relative improvements over HB with 746% and 755%, respectively. Similarly, on answering rate, M2Q-FT5 obtains the highest improvement with 1,727%, while for M2Q-T5 the improvement is 1,478%. This result shows that the buyer's messages in their original form are unsuitable for QA.

M2Q-HYBRID achieves the highest relative improvement across all metrics. This validates our intuition that for shorter messages already in question form, reformulation is not necessary, and in such cases an extractive method provides accurate answers.

Finally, in terms of answer relevance, we see a relative improvement of 2,220% over the extractive baseline, and see no difference between M2Q-FT5 and M2Q-HYBRID-FT5. This shows that not only do we increase answer rates, but also answer precision, as the relative increases in relevance from M2Q are much higher compared to other baselines.

6 Online Deployment and Evaluation

We deployed the more cost-effective M2Q-HYBRID-T5 model, exhibiting a performance equal to M2Q-HYBRID-FT5 according to offline results.

We assess user satisfaction and purchase metrics from millions of e-commerce customers. We split users into two cohorts: a control group (C) whose messages are manually answered by sellers, and a treatment group (T), whose questions are answered by M2Q-HYBRID-T5. We also consider T_{pos}, a subset of T, who provide explicit positive feedback on M2Q answers. We consider the following online evaluation metrics:

- **Purchase Rate – PR**: The ratio of *unique users* who ask a question about a product and buy it within a week, to the total number of users asking a question.
- **Successful Answer Rate – SAR**: The proportion of messages that received an instant answer where buyers were satisfied and did not send it to the seller.†

6.1 Results

Due to confidentiality, the results are reported as relative improvements over the control cohort (C). Table 3 shows the results for the PR and SAR rate metrics.

† We assess the proportion of questions answered when the QA confidence surpasses a threshold.
† If unsatisfied with an instant answer, users can forward their question to the seller.

Purchase Rate – PR: Both treatment groups have significant increases in PR. T_{pos} obtains the highest purchase rate. This is intuitive given that the instant answers are explicitly marked as helpful. This result demonstrates that providing instant answers accelerates customer purchase decisions.

Successful Answer Rate – SAR: On T cohort, users submit significantly fewer messages to sellers (50% relative increase of SAR). For T_{pos}, SAR increases by 276%.

	Control	T	T_{pos}
PR	0.0	+28.57%	+50.88%
SAR	0.0	+57.14%	+276.73%

Table 3: Online evaluation results with real customers. The reported metrics represent relative improvement over the control cohort (C).

7 Conclusion

We proposed M2Q, an approach for automatically answering messages that are sent from buyers to sellers. Offline experiments validated our approach, and live deployment demonstrated that it improves the shopping experience for both buyers and sellers.

Our method efficiently reformulates messages into concise, salient questions optimal for understanding and response by a federated QA system, providing instant answers to buyers. The instant answers feature significantly influences both buyers and sellers, evidenced by a reduction of up to 276% in buyer-to-seller messages. This decrease likely reflects users' satisfaction with the instant responses from M2Q, contributing to enhanced buyer experiences and decreased seller overhead. An empirical online study involving real e-commerce users demonstrated a substantial relative increase in purchase rates by 57.14% when compared to a control group not utilizing M2Q instant answers.

References

1. Ahearne, M., Atefi, Y., Lam, S.K., Pourmasoudi, M.: The future of buyer–seller interactions: A conceptual framework and research agenda. Journal of the Academy of Marketing Science pp. 1–24 (2022)
2. Cao, Y., Li, D., Fang, M., Zhou, T., Gao, J., Zhan, Y., Tao, D.: Tasa: Deceiving question answering models by twin answer sentences attack. arXiv preprint arXiv:2210.15221 (2022)
3. Chen, M., Liu, R., Shen, L., Yuan, S., Zhou, J., Wu, Y., He, X., Zhou, B.: The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 459–466 (2020)
4. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/

5. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V.Y., Huang, Y., Dai, A.M., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models. CoRR **abs/2210.11416** (2022). https://doi.org/10.48550/arXiv.2210.11416, https://doi.org/10.48550/arXiv.2210.11416
6. Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., Zhou, M.: Superagent: A customer service chatbot for e-commerce websites. In: Proceedings of ACL 2017, system demonstrations. pp. 97–102 (2017)
7. Deng, Y., Zhang, W., Yu, Q., Lam, W.: Product question answering in e-commerce: A survey. arXiv preprint arXiv:2302.08092 (2023)
8. Do, X.L., Zou, B., Pan, L., Chen, N., Joty, S., Aw, A.: Cohs-cqg: Context and history selection for conversational question generation. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 580–591 (2022)
9. Faustini, P., Chen, Z., Fetahu, B., Rokhlenko, O., Malmasi, S.: Answering unanswered questions through semantic reformulations in spoken QA. In: Sitaram, S., Klebanov, B.B., Williams, J.D. (eds.) Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 729–743. Association for Computational Linguistics (2023), https://aclanthology.org/2023.acl-industry.70
10. Feng, X., Feng, X., Qin, B.: A survey on dialogue summarization: Recent advances and new frontiers. arXiv preprint arXiv:2107.03175 (2021)
11. Ferguson, N., Guillou, L., Nuamah, K., Bundy, A.: Investigating the use of paraphrase generation for question reformulation in the frank qa system. arXiv preprint arXiv:2206.02737 (2022)
12. Gao, S., Chen, X., Ren, Z., Zhao, D., Yan, R.: Meaningful answer generation of e-commerce question-answering. ACM Transactions on Information Systems (TOIS) **39**(2), 1–26 (2021)
13. Kumar, G., Henderson, M., Chan, S., Nguyen, H., Ngoo, L.: Question-answer selection in user to user marketplace conversations. In: 9th International Workshop on Spoken Dialogue System Technology. pp. 397–403. Springer (2019)
14. Li, Y., Miao, Q., Geng, J., Alt, C., Schwarzenberg, R., Hennig, L., Hu, C., Xu, F.: Question answering for technical customer support. In: Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7. pp. 3–15. Springer (2018)
15. Liao, L.Y., Fares, T.: A practical 2-step approach to assist enterprise question-answering live chat. In: Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 457–468 (2021)
16. Lyu, Q., Zhang, H., Sulem, E., Roth, D.: Zero-shot event extraction via transfer learning: Challenges and insights. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 322–332 (2021)
17. Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., Chen, W.: Generation-augmented retrieval for open-domain question answering. arXiv preprint arXiv:2009.08553 (2020)
18. Masterov, D.V., Mayer, U.F., Tadelis, S.: Canary in the e-commerce coal mine: Detecting and predicting poor experiences using buyer-to-seller messages. In: Proceedings of the Sixteenth ACM Conference on Economics and Computation. pp. 81–93 (2015)
19. McDonald, T., Tsan, B., Saini, A., Ordonez, J., Gutierrez, L., Nguyen, P., Mason, B., Ng, B.: Detect, retrieve, comprehend: A flexible framework for zero-shot document-level question answering. arXiv preprint arXiv:2210.01959 (2022)

20. Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al.: Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:2302.12813 (2023)
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
22. Rennard, V., Shang, G., Hunter, J., Vazirgiannis, M.: Abstractive meeting summarization: A survey. arXiv preprint arXiv:2208.04163 (2022)
23. Samarakoon, L., Kumarawadu, S., Pulasinghe, K.: Automated question answering for customer helpdesk applications. In: 2011 6th International Conference on Industrial and Information Systems. pp. 328–333. IEEE (2011)
24. Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N., Zhou, D.: Large language models can be easily distracted by irrelevant context. arXiv preprint arXiv:2302.00093 (2023)
25. Vakulenko, S., Longpre, S., Tu, Z., Anantha, R.: Question rewriting for conversational question answering. In: Proceedings of the 14th ACM international conference on web search and data mining. pp. 355–363 (2021)
26. Yang, A., Liu, K., Liu, J., Lyu, Y., Li, S.: Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. In: Choi, E., Seo, M., Chen, D., Jia, R., Berant, J. (eds.) Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018. pp. 98–104. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-2611, https://aclanthology.org/W18-2611/
27. Zaib, M., Zhang, W.E., Sheng, Q.Z., Mahmood, A., Zhang, Y.: Conversational question answering: A survey. Knowledge and Information Systems **64**(12), 3151–3195 (2022)