Unlocking Markets: A Multilingual Benchmark to Cross-Market Question Answering

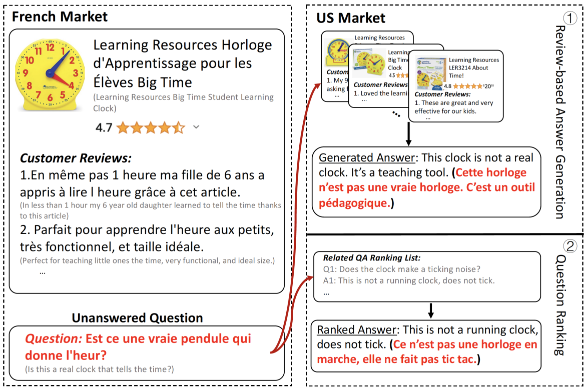
Anonymous ACL submission

Abstract

Users post numerous product-related questions on e-commerce platforms, affecting their purchase decisions. Product-related question answering (PQA) entails utilizing product-related resources to provide precise responses to users. We propose a novel task of Multilingual Cross-market Product-based Question Answering (MCPQA) and define the task as providing answers to product-related questions in a main marketplace by utilizing information from another resource-rich auxiliary marketplace in a multilingual context. To facilitate the research, we propose a large-scale dataset named McMarket, with over 2 million questions across 13 marketplaces in 8 languages. We focus on two subtasks: review-based answer generation and product-related question ranking. Answers are obtained either by generating or ranking from product-related resources (e.g., reviews, questions). For each subtask, we label a subset of McMarket using an LLM and further evaluate the quality of the annotations via human assessment. We then conduct experiments to benchmark our dataset, using a range of models ranging from traditional lexical models to LLMs in both single-market and cross-market scenarios across two datasets. Results show that incorporating cross-market information significantly enhances performance in both tasks.[1]

1 Introduction

Online shoppers on e-commerce platforms post numerous questions to specific products every day (McAuley and Yang, 2015). Since most of these questions remain unanswered, Product-related question answering (PQA) involves providing accurate responses to them. By utilizing product-related information like reviews and metadata, responses to product-related questions can be enriched, offering enhanced depth and authenticity for potential customers (Gupta et al., 2019).


French Market
Learning Resources Horloge d'Apprentissage pour les Élèves Big Time
(Learning Resources Big Time Student Learning Clock)
4.7
Customer Reviews:
Unanswered Question
US Market
①
Review-based Answer Generation
Learning Resources
Learning Resources Big Time Clock
4.3
Custom
1. My 9
asking f
Customer Reviews
1. Loved the learni
Learning Resources LER3214 About Time!
4.8
$20
Customer Reviews:
②
Question Ranking
Related QA Ranking List:

1.En même pas 1 heure ma fille de 6 ans a appris à lire l heure grâce à cet article.
(In less than 1 hour my 6 year old daughter learned to tell the time thanks to this article)
2. Parfait pour apprendre l'heure aux petits, très fonctionnel, et taille idéale.
(Perfect for teaching little ones the time, very functional, and ideal size.)
...

Question: Est ce une vraie pendule qui donne l'heur?
(Is this a real clock that tells the time?)

1. These are great and very effective for our kids.
...

Generated Answer: This clock is not a real clock. It's a teaching tool. (**Cette horloge n'est pas une vraie horloge. C'est un outil pédagogique.**)

Q1: Does the clock make a ticking noise?
A1: This is not a running clock, does not tick.
...

Ranked Answer: This is not a running clock, does not tick. (**Ce n'est pas une horloge en marche, elle ne fait pas tic tac.**)

Figure 1: An example of enhancing product-related QA using cross-market data. ① depicts generating answers with cross-market reviews. ② depicts ranking-related cross-market questions to find the answer.

The recent success in cross-market PQA underscores the capability to effectively leverage relevant questions from a resource-rich marketplace to address questions in a resource-scarce marketplace (Shen et al., 2023; Ghasemi et al., 2023). In this work, we extend the hypothesis that using knowledge from popular marketplaces can improve answers in less common marketplaces, even in other languages. As shown in Figure 1, for a question to a product in the French marketplace (denoted as **main marketplace**) asking if the clock is a real one, we can either address it by examining reviews of the same product or similar ones in the much larger US marketplace (denoted as **auxiliary marketplace**), or ranking related questions from both main and auxiliary marketplaces to find the answer. These multilingual reviews and related questions serve as valuable hints, by saying "it's not a real clock," thereby providing crucial information for the pertinent question at hand.

We, therefore, propose a novel task of *Multilingual Cross-market Product-based Question Answering* (MCPQA). We define this task as *generating the answer to a product-related question in an original marketplace, using information sourced from an auxiliary marketplace with richer*

[1] The code and dataset will be released after paper acceptance. We attach some dataset samples with the submission.

resources, within a multilingual setting. To this end, our initial goal is to address the following research question **RQ1**: *In a multilingual context, how can we utilize an auxiliary marketplace to enhance question-answering in the main marketplace by leveraging product-related resources (i.e., questions, reviews)?* To address **RQ1**, we propose the first large-scale MCPQA dataset, named McMarket, covering 13 different marketplaces (including the **us** auxiliary marketplace and 12 main marketplaces) across 8 different languages. To construct the dataset, we gather data from an existing Amazon product dataset and supplement it with information from user-generated Amazon product question-answering sources. In particular, we provide diverse product information in McMarket, exploring the possible answers using both questions and reviews. In total, our dataset consists of over 2 million product-related questions and 7.7 million product reviews. With McMarket, we then perform comprehensive data analysis to address **RQ1**. We demonstrate a notable increase in the percentage of review-answerable questions across all marketplaces, with support from the auxiliary **us** marketplace.

Given the recent success of large language models (LLMs) in NLP tasks (Touvron et al., 2023a; OpenAI, 2023), their potential application to the MCPQA task prompts our second research question **RQ2**: *Can LLMs benefit the dataset construction in the MCPQA task?* Addressing **RQ2**, we randomly select some questions from McMarket and perform GPT-4 auto-labeling. Specifically, we focus on two widely-studied PQA subtasks under the multilingual cross-market settings, including **review-based answer generation (AG)** (Gao et al., 2019; Chen et al., 2019) and **product-related question ranking (QR)** (Rozen et al., 2021). For AG, we prompt LLMs to judge whether a question can be answered from associated reviews and provide its corresponding answer. This subset is denoted as McMarket$_a$. For QR, given two question answering pairs, we ask LLMs to judge if one helps answer the other and denote the subset as McMarket$_q$. With the two subsets, we then conduct human assessment, scrutinizing the LLM-generated results from multiple angles to ensure their quality meets the required standards. Notably, in McMarket$_a$, 61.8% LLM-generated answers are assumed 'better' than the human ground truth.

Finally, we are interested in answering the research question **RQ3**: *How do existing multilingual and monolingual methods perform in the single- and cross-market scenarios?* To this end, we perform experiments of models on AG and QR subtasks. For each task, we report the performance of state-of-the-art methods under single- and cross-market scenarios on both McMarket and the corresponding subset. We benchmark methods ranging from traditional lexical models (*i.e.,* BM25) to LLM-based approaches (*i.e.,* LLaMA-2, Flan-T5). We demonstrate the superiority of cross-market methods against their single-market counterparts.

In conclusion, our contributions are as follows:

- We propose a novel task named MCPQA, where we leverage product-related information from an auxiliary resource-rich marketplace to answer questions in a resource-scarce one in a multilingual setting. Specifically, we investigate two subtasks named AG and QR.
- We benchmark a large-scale real-world dataset named McMarket to facilitate the research in the MCPQA task. We also collect two LLM-annotated subsets and adopt human assessment to ensure the dataset's quality.
- To provide a comprehensive evaluation of the task and verify the superiority of cross-market methods, experiments are performed under both single/cross-market scenarios.

2 Related Work

Product-related QA. Product-related QA (PQA) seeks to address consumers' general inquiries by utilizing diverse product-related resources such as customer reviews, or the pre-existing QA sections available on a retail platform (Yu et al., 2012; Deng et al., 2023). Among the existing literature in this area, retrieval-based methods have been a popular direction that retrieve related reviews for providing the right answer (Wan and McAuley, 2016; Zhang et al., 2019b, 2020b,a; Yu and Lam, 2018). For example, McAuley and Yang (2015) propose a model that leverages questions from previous records for selecting the relevant review for the question. While most of these works assume there are no user-written answers available, Zhang et al. (2020b) rank answers for the given question with review as an auxiliary input. Another line of research (Gao et al., 2019; Chen et al., 2019; Gao et al., 2021; Feng et al., 2021) investigates answer generation grounding on retrieved product-related documents. More recently, Ghasemi et al. (2023) introduce a novel task of utilizing available data in a resource-rich marketplace to answer questions in a resource-scarce marketplace. Building upon their research, we explore multilingual contexts, examining marketplaces with non-English content.

Cross-domain and cross-lingual QA. Our work can be seen as a special format of cross-domain QA, which involves addressing questions that span different domains or fields of knowledge (Qu et al., 2020; Liu et al., 2019; Longpre et al., 2020). For instance, Yu et al. (2017) propose a general framework that effectively applies the shared knowledge from a domain with abundant resources to a domain with limited resources. Also, cross-domain QA is often with a close connection to cross-lingual QA in the sense that both involve transferring knowledge and understanding from one domain or language to another. (Artetxe et al., 2019; Clark et al., 2020; Zhang et al., 2019a). Asai et al. (2020) expand the scope of open-retrieval question answering to a cross-lingual setting, allowing questions in one language to be answered using contents from another language. Recently, Shen et al. (2023) introduce a multilingual PQA dataset called xPQA where cross-market information is also leveraged to aid the product-based question answering. Compared to these datasets, more diverse information is provided in McMarket, exploring the possible answers with both questions and reviews available.

3 Problem Formulation

We investigate two subtasks of the MCPQA task, *review-based answer generation (AG)* and *product-related question ranking (QR)*, where answers to a product question are obtained by a generative or ranking way, respectively.

AG. In this task, we assume that the answer can be obtained from the reviews of the product (or similar products). Based on the setting in Gupta et al. (2019), we define this task in a multilingual cross-market scenario. Given a question Q in the main marketplace M_T, we first retrieve and rank all the related reviews from similar items within both M_T and auxiliary marketplace M_A. Given the retrieved review set $\Omega = \{R_1, ..., R_k\}$, we predict if Q is answerable from it by assigning a binary label t. If yes, a generative function Γ is learned: $A = \Gamma(Q, \Omega)$, so that answer A is generated with Q and Ω as input.

QR. Following the problem setting in Ghasemi et al. (2023), we assume that there are similar questions already asked about the product or similar products in other marketplaces. Therefore, given a *main marketplace* in language L_M, denoted as M_T, which usually suffers from resource scarcity of the number of knowledgeable users answers, M_T consists of several items $\{I_1, ..., I_m\}$, where each I_k contains a set of question answering pairs $\{QA_{k1}, ...QA_{kn}\}$. Besides, there also exists a high-resource marketplace M_A, denoted as the *auxiliary marketplace* (the **us** marketplace in our case) in language L_A (note that in some cases L_A can be the same as L_M). Similarly, M_A also includes several items $\{I'_1, ..., I'_z\}$, where we can assume $z >> m$. The task is defined as, for a given question Q in the main marketplace M_T, in a multilingual setting, we rank the questions from both M_T and M_A to take the corresponding answers of the top ranks as the possible answer to Q.

4 Data Collection & Analysis

We describe how we collect McMarket (see pipeline in Appendix A) and perform several analysis to answer **RQ1** and **RQ2**.

4.1 Data collection

4.1.1 Data preprocessing

We construct our dataset on top of an Amazon product dataset called XMarket (Bonab et al., 2021). XMarket includes authentic Amazon product metadata and user-generated reviews. Specifically, we sample 13 marketplaces covering 8 different languages from the XMarket Electronic category, including 12 as main marketplaces and the additional **us** marketplace as the auxiliary marketplace. For each marketplace, we gather metadata and reviews for each product from XMarket. We also collect the question-answering pairs posed by the users by crawling the Amazon website. We then provide the corresponding English translation for the non-English contents to ensure help users fully understand and evaluate the data. Specifically, we adopt a widely-used professional translation tool named DeepL Pro[2] for all the question-answer translation and the pre-trained NLLB model (team et al., 2022) fine-tuned on each non-English language for review translation. We ensure the translation quality and

[2] https://www.deepl.com/

Name	Market	Language	Product	Question	Review	Avg. Question per Market
xPQA (Shen et al., 2023)	12	12	16,615	18,000	-	1500
XMarket-QA (Ghasemi et al., 2023)	2	1	34,100	4,821,332	-	2,410,666
semiPQA (Shen et al., 2022)	1	1	-	11,243	-	11,243
SubjQA (Bjerva et al., 2020)	1	1	-	10,098	10,098	10,098
ReviewRC (Xu et al., 2019)	1	1	-	2,596	959	2,596
AmazonQA (Gupta et al., 2019)	1	1	155,375	923,685	8,556,569	923,685
Amazon (McAuley and Yang, 2015)	1	1	191,185	1,447,173	13,498,681	1,447,173
McMarket	13	8	30,724	2,700,179	7,706,519	207,706

Table 1: Comparison of McMarket with existing PQA datasets. The detailed statistics are listed in Appendix E.

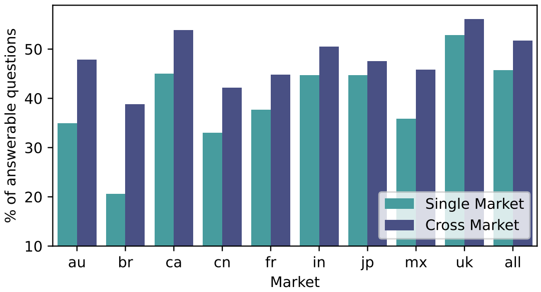


Figure 2: Portion of answerable questions using single/cross-market review information.

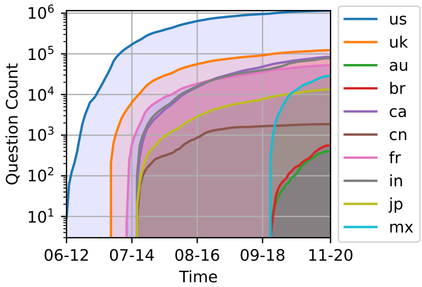


Figure 3: Temporal gap analysis.

provide a detailed assessment in Appendix B. To the best of our knowledge, this is the first multilingual cross-market QA dataset with questions and reviews in the community (license see Section 7).

4.1.2 LLM annotation

For the two concerned subtasks, we both provide LLM-labeled data for supervised training. Specifically, we randomly select a small portion of data from McMarket and instruct GPT-4 to perform annotation. For AG, we randomly select 1000 questions per marketplace.[3] Then, we follow the typical top-K pooling technique (González and Gómez, 2007) and pool the top five retrieved reviews from a variety of retrieval methods. Next, we instruct GPT-4 to evaluate whether the question is answerable. If yes, it generates an answer with the question and reviews as input. If no, it is instructed to output the reason and 'no answer'. We denote this subset as McMarket$_a$. For QR, we randomly select 200 questions from each marketplace. Employing the same strategy, we retrieve the top five related question-answering pairs from both the main and auxiliary marketplaces. Consequently, we acquire 1,000 question-answering pairs for each marketplace, with 9k pairs in total. Then, GPT-4 is instructed to determine if the retrieved QA pairs would be useful in answering the original question by assigning a score from 0–2, representing '*Very useful*', '*Partially useful*', and '*Not useful*', respectively. We denote this subset as McMarket$_q$. For simplicity, we perform GPT-4 labeling on translated English contents. Details see Appendix C.

4.2 Data analysis

4.2.1 Dataset overview

Overall, McMarket covers 13 different marketplaces and 8 languages, ranging from marketplaces with a small scale (*i.e.,* **au**, **br**) to marketplaces with rich resources (*i.e.,* **uk**, **us**). It contains over 2 million product-related questions, 7 million reviews, and 30k unique products in total.

We compare McMarket with existing PQA datasets. According to Table 1, McMarket exhibits advantages in various aspects: (1) **contains multiple languages** – we provide product, question, and review information in the original text of their respective marketplaces and additionally offer the corresponding English translations; (2) **supports cross-market QA** – our dataset is designed to facilitate question answering research across different marketplaces, enhancing its utility for cross-market analyses and evaluations; (3) **includes diverse information** – compared with existing multilingual PQA dataset, McMarket encompasses comprehensive question and review information within a cross-market setting, paving the way for more diverse research tasks in the future; (4) **is large in scale** – overall, McMarket surpasses most PQA datasets in

[3] For the **au** marketplace, the total is 584 questions, so we sample all of them.

	Very Bad	Bad	Good	Very Good
Correctness	2.5	0.9	8.5	88.1
Completeness	4.9	1.3	15.6	78.2
Relevance	3.5	2.7	13.4	80.4
Naturalness	0.8	0.9	5.4	92.9
Better than Ground Truth				**61.8**

Table 2: Human evaluation on McMarket$_a$. All the numbers are shown in percentage.

	Incorrect	Partially correct	Correct
Portion	6.0	10.9	83.0
Overall Precision			98.2
Overall Recall			97.4
Overall F1			**97.6**

Table 3: Human evaluation on McMarket$_q$. All the numbers are shown in percentage.

terms of size, ensuring it comprises a substantial amount of data for experimentation and analysis.

4.2.2 Cross-market QA analysis

To answer **RQ1**, we compare the effect of product-related resources (*i.e.,* reviews) on question answering under both single- and cross-market scenarios. Figure 2 shows the comparison of answerable questions based on both single- and cross-market retrieved reviews in McMarket.[4] We notice that the portion of answerable questions gets raised in every marketplace with cross-market reviews, with a particularly significant uplift observed in low-resource marketplaces (*i.e.,* **br**). This verifies the transferability of knowledge across marketplaces and underscores the advantages of leveraging cross-market information in enhancing the performance of product QA models.

We further analyze the temporal characteristics of McMarket. Figure 3 illustrates the cumulative sum of the number of QA data available on all the items in all marketplaces. There are several notable observations: 1) at the beginning, all marketplaces feature very few QA data. 2) At each timestep, the most resource-rich marketplace (*i.e.,* **us**) always dominates the number of QA data compared to other marketplaces. 3) Over time, the resource intensity levels of different marketplaces continue to change. For example, the number of QA data in **mx** surpasses that in **cn** and **jp** after 2018/09. We further observe that, on average, over 70% of the questions in the main marketplace have already been answered in the **us** auxiliary marketplace for the same item, even before the first question in the main marketplace receives an answer. These findings confirm the practicality and importance of exploring how auxiliary marketplaces can be utilized as valuable resources for PQA.

4.2.3 LLM-generated data analysis

To assess the quality of LLM-generated data, we perform several analyses. On both McMarket$_a$ and McMarket$_q$, we randomly sample 500 questions with the average of 50 questions from each marketplace, and hire 3 crowd-workers for each task[5] and instruct them to manually assess the GPT-4 labels (details see Appendix F).

AG. For McMarket$_a$, we ask the crowd-workers to assess GPT-4-generated answers in terms of correctness, completeness, relevance, and naturalness. For each metric, we asked them to assign a score from -2 to $+2$ to assess the answer quality, with -2 representing 'very bad' and $+2$ representing 'very good.' We also asked them to choose the better answer between the GPT-4 and human-provided answers. They were also asked to provide their reasons without knowing the true category, mitigating bias towards longer and more detailed responses. We note a high agreement among annotators, with a 0.76 IAA score. From Table 2, we note that GPT-4 answers demonstrate reasonable performance in terms of every metric. Surprisingly, our findings reveal that in the majority of cases, human assessors perceive GPT-4 results to be better than human-generated ground truth. It is worth noting that GPT-4's outcomes are derived directly from review information, whereas human ground truth relies on both reviews and actual user experiences.

QR. For McMarket$_q$, we ask the crowd-workers to judge the quality of the question ranking generated by GPT-4, by assigning a score between 0–2 to each sample, where 0 denotes GPT-4 answers are not correct, 1 as partially correct, and 2 as completely correct. Furthermore, we instruct the annotators to provide their own judgment of the ranking score if they mark GPT-4 answers as 0 or 1. We also observe high agreement in this task with the IAA score 0.83. Table 3 shows that the quality of the generated question ranking results by GPT-4 is also deemed satisfactory, achieving over 93% correctness in question ranking pairs and an overall F1 score of 97.6%.

[4] We adopt the answerable question prediction model in Gupta et al. (2019) to predict if a question is answerable or not given the review information.

[5] We hire the crowd-workers via a professional data management company named Appen (https://appen.com/).

	Method	au		br		ca		cn		fr		in		jp		mx		uk		AVG	
		B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R
Single	BM25	6.1	7.0	4.9	6.9	6.9	7.7	4.8	5.2	8.0	8.1	4.7	5.9	11.0	9.6	7.0	8.2	10.3	9.3	8.0	7.9
	BERT	7.4	7.3	9.0	5.3	7.3	6.8	5.4	5.0	8.5	7.2	5.1	4.8	10.6	8.7	9.4	7.7	9.5	8.2	7.9	7.0
	T5	15.5	11.4	14.3	12.6	16.4	12.1	13.5	10.7	16.5	11.5	12.8	9.9	22.6	15.6	20.2	14.4	18.9	13.3	16.9	12.2
	mT5	6.2	5.3	8.1	9.2	14.3	10.0	19.5	11.8	15.5	10.7	9.7	8.7	26.3	13.3	12.2	9.4	14.6	9.6	13.7	9.7
	Llama-2*	10.2	14.7	16.4	17.1	15.9	13.1	14.8	13.6	18.3	14.2	13.5	13.1	26.6	19.7	22.3	16.6	20.1	18.3	17.8	15.4
Cross	BM25	10.6	7.9	9.0	6.1	7.8	7.9	4.6	5.4	9.0	8.2	5.6	6.1	11.3	9.5	9.9	9.1	10.4	9.2	8.9	8.0
	BERT	10.5	8.1	9.5	6.4	8.5	8.9	5.8	5.1	9.8	8.3	6.1	7.3	11.8	9.6	10.4	8.7	11.4	10.3	9.4	9.0
	Exact-T5	14.0	11.8	16.6	13.0	18.2	11.9	13.0	11.0	18.1	11.3	12.5	10.1	22.7	15.0	20.3	14.2	20.6	13.7	17.9	12.3
	T5	16.1	11.3	17.0	14.1	17.0	12.7	15.1	11.3	19.4	12.6	13.2	10.6	23.6	16.0	22.3	16.6	20.2	15.4	18.1	13.5
	Exact-Llama-2*	19.5	15.1	17.4	15.5	16.4	13.8	15.6	11.4	21.6	17.6	**16.9**	**15.1**	27.3	17.8	24.7	17.8	22.4	19.8	20.1	17.0
	Llama-2*	**21.4**	**20.6**	**18.9**	**19.5**	**19.5**	**14.4**	**17.6**	**15.5**	**22.0**	**19.0**	16.5	15.0	**29.5**	**18.6**	**25.7**	**19.2**	**25.0**	**22.7**	**21.7**	**18.3**

Table 4: Experimental results of AG on McMarket with human-provided answers as ground truth. * denotes LLM-based methods. The best-performing model in the single-market setting is in light grey, while models in dark grey are distinguished from their Exact-counterparts. All bold numbers pass the significance t-test at 0.05 level.

	Method	au		br		ca		cn		fr		in		jp		mx		uk		AVG	
		B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R	B	R
Single	BM25	10.3	11.7	10.7	12.5	8.3	13.0	8.5	10.1	11.6	15.7	11.7	14.3	12.8	12.1	13.3	13.6	12.4	14.7	10.7	13.3
	BERT	12.4	10.0	14.8	8.7	11.3	8.8	8.5	7.1	11.1	10.2	12.0	10.6	10.9	9.0	14.1	9.5	9.0	11.1	10.8	9.5
	T5	29.8	27.0	26.7	33.6	29.2	27.4	31.1	24.2	34.9	30.8	29.0	32.2	31.1	27.0	27.2	26.5	29.5	25.9	29.9	28.4
	mT5	10.6	14.3	5.2	13.5	6.8	10.4	41.1	26.4	19.9	17.4	9.2	14.7	34.2	29.1	24.5	16.3	7.2	13.5	18.0	17.4
	Llama-2*	35.7	34.3	37.6	**40.8**	36.3	37.2	38.7	34.3	35.7	32.6	34.4	35.8	34.7	32.4	35.9	34.7	35.4	37.0	35.4	35.9
Cross	BM25	13.5	11.0	12.9	10.0	13.4	12.2	7.4	8.5	12.8	13.0	14.6	15.0	11.6	10.1	15.5	12.6	12.0	15.2	12.6	12.0
	BERT	15.8	10.6	15.7	11.0	14.4	9.8	6.8	8.1	12.2	14.2	13.0	12.1	13.8	11.3	15.7	11.1	10.1	13.1	12.9	11.3
	Exact-T5	30.9	28.2	30.1	29.0	29.3	30.7	29.8	26.7	34.7	31.7	31.8	30.3	30.0	24.6	27.3	28.0	29.1	25.9	30.3	28.4
	T5	32.0	30.2	31.0	28.6	29.9	29.7	32.1	26.8	32.2	31.5	30.1	32.4	36.3	29.9	29.4	27.6	30.2	26.0	31.4	29.1
	Exact-Llama-2*	**37.0**	34.6	34.1	32.6	38.0	39.9	33.0	35.2	**40.8**	**44.3**	36.2	40.2	38.0	34.7	38.4	37.8	35.2	37.9	36.7	37.3
	Llama-2*	35.9	**37.4**	**38.0**	37.9	**39.2**	**40.2**	**39.1**	**36.9**	39.6	41.7	**37.0**	**41.0**	**40.9**	**35.2**	**38.8**	**37.1**	**35.9**	**38.5**	**38.4**	**38.5**

Table 5: Experimental results of AG on McMarket$_a$, where LLM-generated answers are adopted as ground-truth.

5 Experiments

5.1 Experimental setup

Dataset. We perform experiments on AG and QR. For each task, we report the single/cross-market results on the whole dataset and its subset.

For AG, on the McMarket dataset, we first adopt the BERT classifier trained in (Gupta et al., 2019). It assesses each question based on the review information, categorizing them as either answerable or unanswerable. Subsequently, we employ it to abandon all unanswerable questions. We then split the training/validation/testing sets following the portion of 70/10/20%, resulting in 183,092/24,973/49,958 samples, respectively. On the McMarket$_a$ dataset, we also split the data into three sets with the same portions. Specifically, we adopt the GPT-4 generated answers as the ground truth. In the single-market setting, we retrieve the top K reviews from the main marketplace before generating the answers[6]. In the cross-market setting, we retrieve the reviews from both the main and auxiliary marketplaces. We report the generation performance of baselines on the testing set.

For QR, we first rank products, then among the top N products, we rank the top K questions[7]. Since McMarket does not come with any ground-truth ranking results, we perform unsupervised training and adopt GPT-4-labeled data, McMarket$_q$, as the testing set. Besides, to further test the performance of supervised methods on this task, we split McMarket$_q$ into three sets, with 1260/180/360 samples in each. We then train each model on the training set and report results on the testing set.

Evaluation metrics. We adopt several evaluation metrics to assess the performance of models on two tasks. For AG, we compare the model-generated answers with ground-truth user answers using BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) scores. For QR, we report major information retrieval metrics, namely, mean reciprocal rank (MRR) and Precision@3 to evaluate the ranking performance of different methods.

Compared methods. For AG, we first directly rank and select a related review as the answer with methods such as BM25 (Robertson and Zaragoza, 2009), BERT (Devlin et al., 2019). Besides, several generative methods such as T5 (Raffel et al., 2019), LLaMA-2 (Touvron et al., 2023b), are leveraged to train the model to generate the answer given the question and reviews. Specifically, under the cross-market scenario, Exact-model means that in the auxiliary marketplace, we only use reviews from the same item before performing answer generation. For QR, on McMarket, we report ranking methods that do not involve any training (*i.e.,* BERT,

[6] We choose K=5 in our case.

[7] Following Ghasemi et al. (2023), we use N=3 and K=50.

	Method	**au**		**br**		**ca**		**cn**		**fr**		**in**		**jp**		**mx**		**uk**		AVG	
		M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P
Single	BM25	24.5	16.9	15.2	18.3	31.5	28.7	22.0	28.7	21.0	34.7	44.4	46.0	23.8	31.5	28.9	38.7	38.4	40.2	27.7	31.5
	BERT	26.9	43.0	18.2	35.0	30.4	42.8	18.2	34.3	17.7	40.8	47.9	52.7	28.5	34.2	30.0	47.0	40.0	51.8	28.6	42.4
	mBERT	25.9	33.0	16.1	26.7	32.7	33.5	18.5	30.0	17.9	31.2	45.2	46.2	24.1	32.5	32.8	40.2	39.9	43.7	28.1	35.2
	UPR-m	30.4	46.0	21.9	39.3	31.9	48.0	36.2	45.5	36.3	43.7	25.7	56.3	34.7	43.3	39.5	54.2	32.5	52.7	32.1	47.7
	UPR-l*	38.9	48.8	27.8	43.3	36.5	49.7	38.1	48.3	42.5	47.3	35.2	59.8	43.3	47.2	49.0	57.2	38.9	55.5	38.9	50.8
Cross	BM25	51.2	45.2	47.4	40.0	51.0	47.5	50.2	46.8	50.8	44.3	58.0	57.5	54.6	45.5	59.0	54.3	50.8	57.5	52.6	48.7
	Exact-BERT	50.7	38.8	49.1	41.8	48.8	47.0	46.2	46.5	50.1	44.7	59.0	57.3	54.8	45.8	59.3	55.7	51.2	57.3	52.1	48.3
	BERT	52.3	45.7	49.7	42.8	50.4	48.8	49.3	44.2	49.4	43.5	60.5	58.3	55.9	46.0	59.7	57.0	52.5	59.3	53.3	49.5
	CMJim	57.5	56.7	52.4	49.3	53.3	57.7	54.0	50.5	56.9	54.3	62.9	66.8	58.4	53.2	64.9	63.8	52.9	62.7	57.0	57.2
	UPR-m	59.1	55.5	**57.8**	56.0	54.3	58.5	52.8	52.1	54.9	52.3	64.1	64.3	57.5	52.9	62.8	63.7	53.6	64.5	57.4	57.8
	Exact-UPR-l*	59.3	56.0	56.3	57.1	**59.7**	59.5	54.4	53.7	55.4	54.0	65.6	68.8	58.5	53.3	62.4	62.9	54.1	62.8	58.4	58.7
	UPR-l*	**60.0**	**59.5**	57.7	**57.5**	59.0	**63.2**	**61.1**	**54.8**	**57.8**	**58.0**	**67.2**	**70.5**	**62.8**	**56.0**	**67.2**	**66.2**	**59.0**	**66.3**	**60.5**	**60.9**

Table 6: Unsupervised experimental results of the QR on McMarket. Where M and P denote MRR and Precision@3, respectively. * denotes LLM-based methods.

	Method	**au**		**br**		**ca**		**cn**		**fr**		**in**		**jp**		**mx**		**uk**		AVG	
		M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P
Single	BERT-f	32.7	44.4	25.8	48.9	30.0	42.2	31.7	35.6	45.8	47.8	46.2	64.4	51.1	48.9	46.4	58.9	54.4	61.1	40.5	50.2
	mBERT-f	32.8	41.1	21.9	40.0	27.5	40.0	29.4	34.4	41.9	45.6	42.9	56.7	48.6	41.1	42.3	51.1	52.9	56.7	37.8	45.2
	T5	29.4	42.2	23.3	41.1	31.7	38.9	31.3	30.9	42.0	45.1	43.8	58.4	49.7	47.8	44.4	54.1	53.9	56.4	38.8	46.1
	monoT5	30.1	44.4	23.1	41.1	31.3	43.2	31.4	31.1	43.2	46.7	49.4	63.3	53.5	49.9	47.8	54.4	53.4	58.9	40.4	48.1
	Flan-T5*	39.7	51.1	26.9	50.0	34.0	46.7	38.3	42.2	52.2	54.4	51.4	63.3	54.8	64.4	49.3	60.0	55.8	62.2	44.7	54.9
Cross	Exact-BERT-f	46.4	45.6	40.0	51.1	51.5	47.8	49.4	45.6	52.3	53.2	49.3	66.0	53.4	47.8	48.9	63.3	58.7	66.7	50.0	54.1
	BERT-f	58.6	54.4	52.3	54.4	55.3	53.3	56.2	46.7	53.9	55.6	65.8	70.0	56.0	52.2	63.2	71.1	59.6	70.0	57.9	58.6
	Exact-monoT5	52.6	48.9	50.7	53.8	54.6	55.6	54.4	44.9	53.2	53.1	63.1	71.0	56.9	52.1	62.8	67.8	59.3	66.8	56.4	57.1
	monoT5	52.9	53.3	51.4	52.2	54.1	56.7	56.8	44.4	52.8	52.2	68.1	75.6	56.8	53.3	62.9	68.9	58.2	67.8	57.1	58.3
	Exact-Flan-T5*	60.8	60.3	55.7	**56.9**	61.3	59.2	57.6	55.2	58.1	57.8	67.2	73.3	57.1	54.3	63.9	74.9	63.0	**73.9**	60.5	62.9
	Flan-T5*	**63.6**	**62.2**	**56.9**	55.6	**62.9**	**61.1**	**59.7**	**57.8**	**60.8**	**61.1**	**69.7**	**76.7**	**60.4**	**56.7**	**64.3**	**75.6**	**63.6**	72.2	**62.4**	**64.3**

Table 7: Supervised experimental results of QR using McMarket$_q$.

UPR (Sachan et al., 2022)) or methods that perform unsupervised training (*i.e.,* CMJim (Ghasemi et al., 2023)). On McMarket$_q$, we adopt supervised fine-tuning methods (*i.e.,* BERT-f/monoT5 (Nogueira et al., 2020)), and report testing performance. We report the performance under random seed 42. More experimental details see Appendix G.

5.2 Experimental results

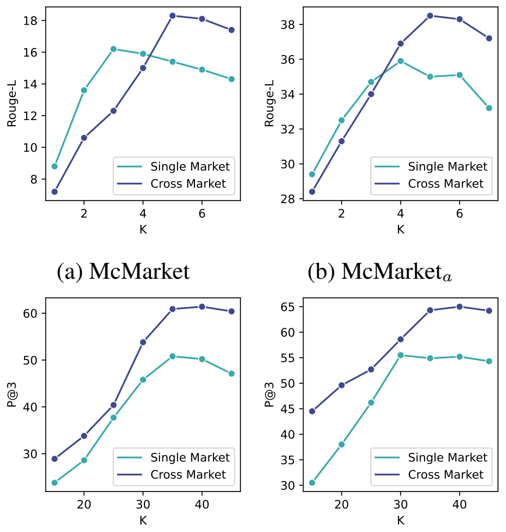
5.2.1 Review-based answer generation

Tables 4 and 5 show the single/cross-market answer generation performance on McMarket and McMarket$_a$ datasets[8]. We have the following observations: first of all, cross-market models have superior overall performance in all marketplaces compared with methods in the single-market setting. This result verifies **RQ1** from the model perspective, showing that external resources (*i.e.,* reviews), from auxiliary marketplaces, can significantly contribute to improved outcomes in the main marketplace. A clear advantage of LLMs over traditional methods is evident across various marketplaces. Notably, LLaMA-2 outperforms the overall cross-market McMarket dataset, with a notable ROUGE improvement from 13.5 in T5 to 18.3. Similarly, in McMarket$_a$, the overall ROUGE score sees significant enhancement, rising from 29.1 to 38.5. This provides an answer for **RQ3**, offering insights into the efficacy and potential advancements of LLMs.

5.2.2 Product-related question ranking

Tables 6 and 7 show the question ranking results within the single/cross-market scenario on two datasets. We notice that most observations from Section 5.2.1 still hold. For example, performance advantages persist in product-related question ranking compared to a single-market scenario. This shows that the large number of relevant questions in the auxiliary marketplaces help address similar questions in a low-resource marketplace. Furthermore, the performance boost is more obvious in marketplaces with a smaller scale (*i.e.,* **au**, **br**) compared with marketplaces with a larger scale (*i.e.,* **uk**). For instance, the P@3 BM25 performance exhibits an improvement 28.3 and 21.7 for **au** and **br** marketplaces, respectively, compared with 17.3 in **uk** on McMarket. We also find that in the cross-market setting, the Exact-models have a weaker overall performance than their original counterparts (*i.e.,* Exact-T5/Llama-2 v.s. T5/Llama-2). For example, on McMarket$_q$, the cross-market Exact-Flan-T5 is 1.4 weaker in terms of overall P@3 compared with Flan-T5. This demonstrates that valuable information can be found within similar products from auxiliary marketplaces, even when they possess slightly different titles. We list some cases in Appendix H to elaborate this.

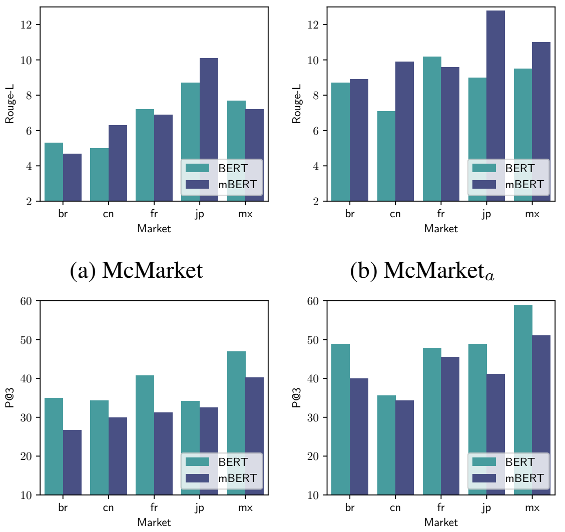
[8] We report performance on 9 marketplaces and leave the 3 untranslated raw marketplaces (**es**, **it**, **de**) for future work.



(a) McMarket (b) McMarket$_a$

(c) McMarket (d) McMarket$_q$

Figure 4: K-value analysis on different marketplaces. The upper row is on AG, the lower is QR.



(a) McMarket (b) McMarket$_a$

(c) McMarket (d) McMarket$_q$

Figure 5: Multilingual analysis on non-English marketplaces. The upper row is on AG, the lower is QR.

6 External Analysis

6.1 Hyperparameter analysis

We investigate the effect the number of retrieved product-related resources (*i.e.,* questions, reviews) K under both single/cross-market scenarios. We report the average performance among every marketplace on both McMarket and the corresponding subset. The results are shown in Figure 4.

We observe that in AG, initially, the performance of Llama-2 in the cross-market setting is inferior to that in the single-market. However, after increasing the value of K, the optimal K value in the cross-market scenario surpasses that in the single-market. This tendency indicates that richer information is contained in the cross-market reviews. In QR, the ranking performance in the single-market scenario begins to decline when K is around 50. This indicates that some less relevant questions are retrieved, negatively impacting the results. Conversely, in the cross-market scenario, as a greater number of relevant questions are accessible, it helps to effectively mitigate this issue.

6.2 Multilingual analysis

We undertake a comparative analysis between translated and non-translated contents to delve deeper into performance variations across non-English marketplaces. In particular, within the single-market scenario, we compare mBERT with BERT in 5 non-English marketplaces. Here, 'mBERT' refers to a setup where all contents and the model itself are preserved and fine-tuned in their original language without translation. The results are shown in Figure 5. We notice that in the AG task, concerning some non-Latin languages (*i.e.,* **cn**, **jp**), the performance of single-market mBERT without translation results in higher score compared with T5 and BERT on two datasets. However, we observe opposite results in some other non-English marketplaces (*i.e.,* **fr**). Besides, in the QR task, the performance of mBERT is inferior to the translated BERT model. This underscores a crucial future direction for this task: effectively enhancing performance in non-English marketplaces, an aspect that has been relatively underexplored.

7 Conclusions

We propose a novel task of Multilingual Cross-market Product-based Question Answering (MCPQA). We hypothesize that product-related information (*i.e.,* reviews/questions) from a resource-rich marketplace can be leveraged to enhance the QA in a resource-scarce marketplace. Specifically, we focus on two different tasks: AG and QR. To facilitate the research, we then propose a large-scale dataset named McMarket, which covers over 2 million questions across 13 marketplaces and 8 languages. We also provide LLM-labeled subsets for the two tasks, namely McMarket$_a$ and McMarket$_q$. We conduct experiments to compare the performance of models under single/cross-market scenarios on both datasets and demonstrate the superiority of cross-market methods in this task.

Limitations

The task of PQA holds significant potential in improving user experiences on e-commerce platforms. However, there are several limitations and challenges associated. One major challenge is the quality and reliability of the information available for answering user questions. Even though we make sure all of the information comes from real user-generated data, the reviews and QA pairs might still contain biased or inaccurate information. Furthermore, language barriers and the availability of data in multiple languages add complexity to the task of product-related QA, particularly in cross-lingual scenarios. We discovered that the performance of non-English content remains unsatisfactory compared to results in English marketplaces. Limited availability of data in low-resource languages further exacerbates this challenge. To address them, continued research and development efforts are still under process which aim at improving data quality, handling language diversity, etc. We discuss it as our future work in Appendix D.

Ethics Statement

Our dataset is derived from the publicly available product question-answering dataset, XMarket (Bonab et al., 2021), which is under the CDLA 1.0 Sharing License and grants academic usage so that follow-up research papers can re-use the data. We adhere to the policies throughout the creation and utilization of this dataset to ensure the protection of user privacy. No personally identifiable information is exposed or utilized in any form during the processes associated with the dataset. Also, we have licensed our data under CC0 1.0 DEED such that it will only be available for academic research purposes to further protect the users. We make sure that individuals sign an agreement stipulating that the dataset will only be used for research purpose when we release the dataset.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*.

Akari Asai, Jungo Kasai, J. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. In *North American Chapter of the Association for Computational Linguistics*.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang Chiew Tan, and Isabelle Augenstein. 2020. Subjqa: A dataset for subjectivity and review comprehension. In *Conference on Empirical Methods in Natural Language Processing*.

Hamed Bonab, Mohammad Aliannejadi, Ali Vardasbi, E. Kanoulas, and James Allan. 2021. Cross-market product recommendation. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019. 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, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 411–419. ACM.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Yang Deng, Wenxuan Zhang, Qian Yu, and Wai Lam. 2023. Product question answering in e-commerce: A survey. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Yue Feng, Zhaochun Ren, Weijie Zhao, Mingming Sun, and Ping Li. 2021. Multi-type textual reasoning for product-aware answer generation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1135–1145. ACM.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2021. Meaningful answer generation of e-commerce question-answering. *ACM Trans. Inf. Syst.*, 39(2):18:1–18:26.

Shen Gao, Zhaochun Ren, Yihong Eric Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 429–437. ACM.

Negin Ghasemi, Mohammad Aliannejadi, Hamed Bonab, E. Kanoulas, Arjen P. de Vries, James Allan, and Djoerd Hiemstra. 2023. Cross-market product-related question answering. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

José Luis Vicedo González and Jaime Gómez. 2007. Trec: Experiment and evaluation in information retrieval. *J. Assoc. Inf. Sci. Technol.*, 58:910–911.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary Chase Lipton. 2019. Amazonqa: A review-based question answering task. In *International Joint Conference on Artificial Intelligence*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. Xqa: A cross-lingual open-domain question answering dataset. In *Annual Meeting of the Association for Computational Linguistics*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Julian McAuley and Alex Yang. 2015. Addressing complex and subjective product-related queries with customer reviews. *Proceedings of the 25th International Conference on World Wide Web*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy J. Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings*.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *North American Chapter of the Association for Computational Linguistics*.

Colin Raffel, Noam M. 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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Ohad Rozen, David Carmel, Avihai Mejer, Vitaly Mirkis, and Yftah Ziser. 2021. Answering product-questions by utilizing questions from other contextually similar products. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 242–253. Association for Computational Linguistics.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen tau Yih, Joëlle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Conference on Empirical Methods in Natural Language Processing*.

Xiaoyu Shen, Akari Asai, Bill Byrne, and A. Gispert. 2023. xpqa: Cross-lingual product question answering in 12 languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, and A. Gispert. 2022. semipqa: A study on product question answering over semi-structured data. *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*.

Nllb team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Mengting Wan and Julian McAuley. 2016. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 489–498.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *North American Chapter of the Association for Computational Linguistics*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Feng Ji, Wei Chu, and Haiqing Chen. 2017. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

Jianxing Yu, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Answering opinion questions on products by exploiting hierarchical organization of consumer reviews. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 391–401. ACL.

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

Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander R. Fabbri, Neha Verma, William Hu, and Dragomir R. Radev. 2019a. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. *ArXiv*, abs/1906.03492.

Shiwei Zhang, Jey Han Lau, Xiuzhen Zhang, Jeffrey Chan, and Cécile Paris. 2019b. Discovering relevant reviews for answering product-related queries. *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1468–1473.

Shiwei Zhang, Xiuzhen Zhang, Jey Han Lau, Jeffrey Chan, and Cécile Paris. 2020a. Less is more: Rejecting unreliable reviews for product question answering. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part III*, volume 12459 of *Lecture Notes in Computer Science*, pages 567–583. Springer.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020b. Answer ranking for product-related questions via multiple semantic relations modeling. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

A Annotation Pipeline

In order to vividly show how we annotate the dataset, we show the annotation pipeline in Figure 6.

B Translation Quality Assessment

In order to ensure the quality of our proposed dataset, we performed some evaluation on the translation accuracy of the DeepL Pro and NLLB translation. For each of the marketplace, we randomly select 100 QA pairs to manually evaluate their correctness, with a mixture of native speakers (**cn** and **jp**) and google translate reference (**fr**, **br** and **mx**). Table 8, 9 show the results. Furthermore, we perform some manual check by comparing the NLLB and DeepL Pro translation. We asked the assessors to check the performance and pick the better one. We see from Table 10 that DeepL Pro has a generally better performance, which explains our motivation of using it for better question translation.

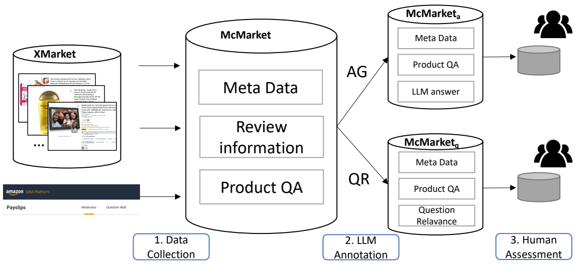


Figure 6: Annotation pipeline of our dataset.

C LLM Annotation Details

To select candidate reviews for LLM answer generation, we combine the matching results by BM25, TF-IDF, Lucene's Classic Similarity, Learning to Rank (LTR), and FastText. We employ GPT-4 as the base LLM to perform automatic annotation. Specifically, `gpt-4-1106-preview` is adopted in our setting. For review-based answer generation, we pass the question, related reviews into the model, and ask GPT-4 to generate if the corresponding answer can be produced from the given information and write the answer if possible. We also instruct GPT-4 to provide the corresponding reason. We use the following prompt:

- *In this task, you will be given a product question, and some reviews. You should judge if the reviews are helpful for answering the question. If yes, please write the corresponding answer and the reason. If no, please give the corresponding reason and provide the answer as no answer. Please output the answer format as: Judgement:yes/no, Reason: , Answer:*

In our setup for product-related question ranking, we follow the annotation setting outlined in Ghasemi et al. (2023). Here, we utilize GPT-4 to evaluate the relevance of other question-answer pairs. The model is presented with two question-answer pairs from distinct products along with their respective product titles. Its task is to assess whether the QA pair associated with the second product proves useful in addressing the questions posed for the first product. Similarly, the model is also requested to provide the reason for making the judgment. The prompt is given as follows:

- *In this task, you will be given two different products, namely, Product A and B, respectively. Each product is associated with a question-answer pair. You should judge if the question-answer pair to Product B is useful for answering the question to Product A. You should assign a score from 0–2, as 0 represents not useful, 1 represents partially useful, and 2 represents very useful. Please also give the corresponding reason for making the decision. Please output the answer format as: Judgement:[score], Reason:*

br	cn	fr	jp	mx
97%	98%	93%	95%	96%

Table 8: Estimated translation accuracy on DeepL Pro.

br	cn	fr	jp	mx
76%	72%	69%	74%	78%

Table 9: Estimated translation accuracy on NLLB.

br	cn	fr	jp	mx
92%	88%	94%	89%	87%

Table 10: Comparison of translation results between NLLB and DeepL, showing the percentage where DeepL outperforms NLLB.

D Future Directions

Future directions for the MCPQA task could involve several areas of exploration. First of all, more efforts could be put in the continued advancement and refinement of multilingual models capable of understanding and generating text across multiple languages. Additionally, three marketplaces (**es**, **de**, **it**) are currently unlabeled, meaning all reviews and question-answer pairs remain in their original, untranslated versions. We are still investigating how models perform when fine-tuned on this untranslated data, particularly in multilingual contexts, and aim to evaluate their question-answering performance accordingly. Based on that, investigation of cross-lingual transfer learning techniques to facilitate knowledge transfer and adaptation between languages could also be a promising direction in this task. This includes exploring approaches for transferring knowledge from high-resource to low-resource languages and vice versa.

E McMarket Statistics

Table 11 shows the detailed statistics of McMarket.

F Human Evaluation Details

We provide a comprehensive annotation guideline to the annotators when they assess the dataset. Specifically, we created a detailed Google document defining each metric, supplemented with 10 sample annotations covering each metric to ensure clarity and understanding. After sharing this document with the annotators, we randomly sample 20 data records for a preliminary annotation phase. These initial annotations undergo a manual check to ensure accuracy and consistency. Once all anno-

	au	**br**	**ca**	**cn**	**fr**	**in**	**jp**	**mx**	**uk**	**us**	**raw**	Total
Language	en	pt	en	cn	fr	en	ja	es	en	en	es, it, de	-
Question Num.	584	1,378	101,126	3,324	66,536	115,829	17,418	34,433	164,848	1,782,092	412,611	2,700,179
Review Num.	3,062	3,650	575,052	1,893	359,703	240,167	130,604	125,317	775,900	4,169,476	1,321,695	7,706,519
Product Num.	85	95	5,432	210	2,199	2,085	903	1,464	4,406	29,976	4,722	30,724
Mean ques. len	12.0±6.6	10.3±6.4	12.7±7.3	10.2±8.1	15.2±6.9	10.1±6.0	20.3±15.7	10.9±6.8	13.6±7.6	13.4±7.8	13.2±6.9	13.3±7.7
Medium ques. len.	10	8	10	8	14	8	14	9	11	11	12	11
Mean review len.	25.5±30.0	17.5±25.3	29.9±50.4	56.4±60.2	39.1±49.7	21.8±42.9	28.7±36.8	28.7±36.8	40.1±68.4	59.3±93.0	47.7±70.1	50.9±81.8
Medium review len.	15	10	14	39	26	10	46	21	20	30	28	26

Table 11: Overall statistics of the McMarket dataset. The length is reported on the token level. For the raw set, we include them in McMarket but leave the discussion to future work in Appendix D.

tations pass this check, the main annotation phase begins. The definition of the human evaluation metrics are listed as follows:

- *Correctness* aims to judge whether GPT-4 answers accurately serve as correct answers to the question, based on the given information. For example, if the question is not answerable from the reviews, GPT-4 should make the corresponding judgment. Otherwise, GPT-4 should first classify the question as answerable, and then give the corresponding answer.
- *Completeness* is designed to determine whether the GPT-4 generated answers are complete and cover all aspects of the question.
- *Relevance* is designed to determine whether the GPT-4 answers are relevant to the question, and whether contain hallucination that does not correspond to the original question.
- *Naturalness* aims to determine whether the GPT-4 answers are smooth and natural. Whether there are obvious language errors and inconsistencies.

G Experimental Details

We implement all of the baselines under the Pytorch framework and the HuggingFace model repository. We conduct all of our experiments using 4 A100 GPUs. For BM25, we use the 'rank-bm25' repository. For all T5-related models, we use 'T5-base' version in Huggingface. For Flan-T5, we use the 'Flan-T5-XL' version. For Llama-2, we use Llama-2-7B and fine-tune it with LoRA adapter. We use all the default parameters in the repository.

To prevent LLM overfitting, we use several strategies including: (1) early stopping (we monitor the model's performance on a validation set and stopping training when performance stops improving to prevent overfitting to the training data), (2) gradient clipping (we limit the magnitude of gradient updates to prevent the model from making large updates that could lead to overfitting), and (3) using some regularization techniques (we use dropout and weight decay to further reduce the risk of overfitting). We provide a detailed explanation of the baseline models we implement.

Review-based answer generation. In this task, we report performance on McMarket and McMarket$_a$. In contrast to utilizing human answers in McMarket, in McMarket$_a$, we employ the GPT-4 generated results as the ground truth. For each dataset, we split the training/validation/testing set with the portion 70/10/20% and report the results on the testing set. The detailed information of each baseline is as follows:

- BM25 (Robertson and Zaragoza, 2009) retrieves the top 5 reviews and adopts the top one directly as the answer.
- BERT (Devlin et al., 2019) adopts a BERT ranker to re-rank the reviews retrieved by the top 100 BM25 results. Then the top 1 review is selected as the answer.
- T5 (Raffel et al., 2019) takes the BM25 top 5 reviews as input and is fine-tuned to generate the corresponding answer.
- mT5 (Xue et al., 2020) is fine-tuned in a similar setting as T5, with the exception that we utilize the mT5 backbone and abstain from using translated non-English contents.
- Exact-T5 (Ghasemi et al., 2023) is an answer generation model based on T5, wherein we initially identify the exact same item in the auxiliary marketplace and exclusively utilize the top 5 reviews among them as input.
- LLaMA-2 (Touvron et al., 2023b) is in a similar setting as T5 but adopts LLaMA-2 as the backbone.

Idx	Question	Single-market ans.	Cross-market ans.	Gold ans.
1	Which Dell Inspiron N4010 model can I buy and be compatible with?	(The express delivery was very fast. It arrived the next day.) I have no idea what to answer.	(I bought this to double my RAM on my new Dell Inspiron 7559.) The Dell Inspiron 7559 model is compatible with this product.	This model is low voltage...
2	Sir is it compatible with x-T3?	(The picture quality is very sharp. Very sharp indeed.) The given review information cannot provide the answer.	(Bokeh is awesome quality of photos is amazing paired with the xt3) Yes, the lens is compatible with xt3.	Yes
3	Will this work for a small fan that's 18 watts?	(Will this work for charding an electric razor?) It provides power out the same as a wall power point. So yes it should charge an electric razor.	(Will a travel fan work with this?) ... A travel fan should work this voltage converter. The converter has a maximum of 200 watts...	I think yes ...
4	Does it have an external microphone input?	(Does it have an audio and hdmi output?) Good morning. It has a micro HDMI output.	(Can you put an external microphone?) No microphone input	Doesn't have it!

Table 12: Real cases of the single/cross-market question answering results. The first two are review-based generated answers. The rest shows answers obtained by product-related question ranking. The information in the bracket shows the retrieved related review/question. All the information is shown in translated English.

- Exact-LLaMA-2 is in a similar setting as Exact-T5 but adopts LLaMA-2 as the backbone.

Product-related question ranking. In this task, we also report results on McMarket and McMarket$_q$. Given that the McMarket$_q$ subset is the only portion in McMarket that contains ranking labels, Table 6 exclusively showcases unsupervised methods that leverage the remaining McMarket as the training set and subsequently present results on the McMarket$_q$ subset. Besides, to show the performance of supervised methods in this task, Table 7 splits McMarket$_q$ as the training/validation/testing set following the same portion as before. Performance is then reported on the testing set.

We first provide details for the unsupervised methods in Table 6:

- BM25 (Robertson and Zaragoza, 2009) reports the top-50 BM25 ranking results.

- BERT (Devlin et al., 2019) performs BERT re-rank on BM25 top results.

- UPR-m (Sachan et al., 2022) is an unsupervised ranking method where we use a PLM to compute the probability of the input question conditioned on a related question. We use T5-base as the backbone.

- UPR-l adopts the same structure as UPR-m but uses T0-3B as the backbone.

- CMJim (Ghasemi et al., 2023) is an unsupervised method that ranks products and their corresponding questions across marketplaces.

- Exact-{BERT/UPR-l} ranks the questions of the item from the main marketplace as well as the exact same item in the auxiliary marketplace.

We then detail the supervised methods in Table 7:

- Bert-f (Devlin et al., 2019) fine-tunes the Bert ranker on the training set.

- T5 is trained to generate the sequence of the ranked questions.

- monoT5 (Nogueira et al., 2020) is another ranking method that takes T5 as backbone. We fine-tune the model on the training set and report the results on the testing portion.

- Flan-T5 (Chung et al., 2022) adopts the same structure as the monoT5 method but replaces the backbone to the Flan-T5-XL LLM.

- Exact-{BERT-f/monoT5/Flan-T5} (Ghasemi et al., 2023) ranks the questions of the item from the main marketplace as well as the exact same item in the auxiliary marketplace.

H Case study

Table 12 demonstrates four real cases concerning single/cross-market question answering. We see that the absence of useful information, such as related reviews or questions, within a single marketplace leads to inaccurate answers. For instance, in

Market	Product title	Question	Reviews	Answer
br	Sony - HDRCX405 HD Video Recording Handycam Camcorder (black)	É compatível com eos 80d?	Objetiva com desempenho muito bom. Estabilização de imagem (IS) funciona muito bem para uso sem tripé. STM com foco silencioso. Cumpre o que promete.	Bom dia, é totalmente compatível.
cn	AKG Pro Audio K612 PRO Over-Ear, Open-Back, Premium Reference Studio Headphones	akg品控真有那么差吗还是一群职业黑?	一言难尽。买了十几天刚煲开右耳时响时不响。现在退货中	没有问题，还可以
fr	ViewSonic VG2439SMH 24 Inch 1080p Ergonomic Monitor with HDMI DisplayPort and VGA for Home and Office, Black	Sur écran webcam il y a t'il du son ? fait t'il webcam et micro en même temps?	Après réception; et déballage : produit simple et mise en marche facile. J'ai commandé deux écrans pour une station de travail. l'utilisateur est à l'aise	Pas le microphone. Webcam ok Son ok
jp	SanDisk Ultra 64GB USB 3.0 OTG Flash Drive With micro USB connector For Android Mobile Devices(SDDD2-064G-G46)	A1954に多用できますか	小さすぎて使いにくい（笑）商品は、ゆうメールですぐに配達されました。	A1954とは、何ですか？キーボードは、英語配列です。
mx	ZOTAC GeForce GT 730 1GB DDR3 PCI Express 2.0 x1 Graphics Card (ZT-71107-10L)	hola, es compatible con Lenovo TS-140?	Excelente producto y buen desempeño. Muy recomendable.	No conozco este equipo, solo se puede instalar en interfaces PIC x16.
uk	Peachtree Audio Deepblue2 High Performance Wireless Bluetooth Music System (Black)	Can you play music through this speaker as a wired device from an mp3 player.	Reluctant to pay so much but couldn't be more happy. Amazing sound quality and worth every penny. You will be blown away.	Yes, with the supplied cable plugged into the headphone jack on the MP3 player and the auxiliary input on the deepblue2.

Table 13: Examples of data samples in McMarket.

Market	Question	Reviews	GPT-4 ans.	Gold Ans.	GPT-4 reason
au	Can you use this with Xbox one?	... These however do not work as an Xbox one gamechat since there is no mic ...	No, you cannot use these headphones with Xbox One as there is no mic	Yes.	Review 1 mentions that these headphones do not work as an Xbox One gamechat as there is no mic.
ca	Will it work with the Nikon D3200	I have the lens fixed on to my Nikon D3200 almost permanently...	Yes, it will work with the Nikon D3200.	Works on my d3100.	The reviews mention that the product is compatible with the Nikon D3200 and they are using it effectively.
in	Does it has diopter correction ring?	Extremely well built. Great lens and use it mostly for bird watching during my hikes...	No answer.	Yes.	None of the reviews mention a diopter correction ring for the product.
jp	Does it come with a conversion plug to connect to an amplifier?	...They don't come with an adapter that can adjust to the standard plug, so be careful.	No, it does not come with a conversion plug to connect to an amplifier.	N/A.	Review 1 states explicitly that the headphones do not come with an adapter to adjust to the standard plug...
mx	How quiet is it?	...Good performance, very quiet; Nice and quiet easily installed and works well...	The product is very quiet.	It is very quiet, reliable, highly recommended	Reviews 2, 3, 4, and 5 directly address the noise level of the product by stating it is 'very quiet' and 'nice and quiet'

Table 14: Examples of data samples in McMarket$_a$. All the data is translated into English.

case 1, the retrieved reviews fail to provide sufficient information, resulting in a generated answer of "I have no idea what to answer." In contrast, relevant and useful information is more likely to be available in the larger auxiliary marketplace. For instance, in case 4, the model successfully retrieves a similar question, "Can you put an external microphone?" from the **us** marketplace, aligning the answer more closely with the ground-truth answer.

I Dataset Examples

We show some examples from McMarket to provide a more comprehensive view of our data. Table 13 shows some examples from McMarket. For each example, we show the title of a product, a random review, and a question-answer pair of the product.

To provide a more comprehensive understanding of our dataset and task, we also show some examples of the GPT-4 annotated McMarket$_a$ (Table 14)

Market	Product A	Product A QA	Product B	Product B QA	tag	GPT-4 reason
au	Neewer 48 Macro LED Ring Flash Bundle with LCD Display Power Control...	Will this work with fuji x-t3 and x-t20? -> As long as they have a hot shoe, it will work. There is several lens ring adaptors for various lens sizes (talking about changeable lenses of course).	Neewer 48 Macro LED Ring Flash Bundle with LCD Display Power Control...	Is this compatible with FujifilmX-T3? -> As long as you have a hotshoe it should work.	2	Both Product A and Product B are the same Neewer 48 Macro LED Ring Flash Bundle, and the questions for both are concerning the compatibility with Fujifilm X-T3...
cn	Kingston Digital Multi-Kit/Mobility Kit 16 GB ...	Hello, what is the writing speed of this micro sdxc? -> Write: 14Mo/s | Read: 20Mo/s ...	Kingston Digital Multi-Kit/Mobility Kit 16 GB...	Speed of the card? -> Class 4 IE 4MB/sec.	1	The answer to Product B provides the class rating of a microSDHC card, though different from Product A...
fr	iPad Air New iPad 9.7 inch 2017 Case...	Good evening, is this case compatible with an iPad 2? Thank you -> Yes, no problem.	iPad Air New iPad 9.7 inch 2017 Case...	Does this case fit the ipad air 2? -> Hi, This case is not compatible with the iPad Air 2.	0	Product A is asking about iPad 2, while Product B is about compatibility with an iPad Air 2...
in	AmazonBasics USB 2.0 ...	Is it compatible with Nintendo switch? -> Dono but working good nice product.	AmazonBasics USB 2.0 ...	Is this compatible with MacOS? -> Yes.	0	The answer to Product B's question does not provide information for A...
uk	HDMI Media Player, Black Mini 1080p Full-HD Ultra...	Is it possible to power this through a usb cable? -> It has to be plugged in using the power lead...	MDN HD1080B 1080p Full-HD Ultra Portable Digital Media Player...	Can it be powered by a USB cable? I see on the pictures that power cable is USB on one end -> The USB port is for an external drive.	2	The question for both Product A and Product B pertains to the power source of the media players and whether they can be powered through a USB cable...

Table 15: Examples of data samples in McMarket$_q$. All the data is translated into English.

and McMarket$_q$ (Table 15), respectively.