Domain Adaptation for Subjective Induction Questions Answering on Products by Adversarial Disentangled Learning

Yufeng Zhang, Jianxing Yu,* Yanghui Rao, Libin Zheng, Qinliang Su, Huaijie Zhu, Jian Yin
School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, 519082, China
Technology Innovation Center for Collaborative Applications of Natural Resources Data in GBA, Ministry of Natural Resources, Guangzhou 510075, China
Key Laboratory of Sustainable Tourism Smart Assessment Technology, Ministry of Culture and Tourism of China
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
Pazhou Lab, Guangzhou, 510330, China
{zhangyf283,yujx26,raoyangh,zhenglb6,suqliang,zhuhuaijie,issjyin}@mail.sysu.edu.cn

Abstract

This paper focuses on answering subjective questions about products. Different from the factoid question with a single answer span, this subjective one involves multiple viewpoints. For example, the question of '*how the phone's battery is?*' not only involves facts of battery capacity but also contains users' opinions on the battery's pros and cons. A good answer should be able to integrate these heterogeneous and even inconsistent viewpoints, which is formalized as a subjective induction QA task. For this task, the data distributions are often imbalanced across different product domains. It is hard for traditional methods to work well without considering the shift of domain patterns. To address this problem, we propose a novel domain-adaptive model. Concretely, for each sample in the source and target domain, we first retrieve answer-related knowledge and represent them independently. To facilitate knowledge transferring, we then disentangle the representations into domain-invariant and domain-specific latent factors. Moreover, we develop an adversarial discriminator with contrastive learning to reduce the impact of out-of-domain bias. Based on learned latent vectors in a target domain, we yield multi-perspective summaries as inductive answers. Experiments on popular datasets show the effectiveness of our method.

1 Introduction

With the popularity of e-commerce platforms, many merchants publish various kinds of content about products on the Web (Khern-am nuai et al., 2023). Based on such a large amount of content, it is difficult for consumers to seek useful knowledge for making informed purchase decisions. To tackle this information overload problem, consumers turn to ask questions on product attributes, functions, and user experiences via the forums. Since it is intractable to manually reply to so many questions, product question-answering systems (*PQA*) have emerged. Among these questions, the factoid ones have been well studied (Feng et al., 2021). For instance, for a factoid question "*What is the operating system of this laptop?*" we can simply extract a span "*MacOS Ventura*" from the input text as the answer. However, many complex questions are still less investigated, such as the subjective ones asking about personal feelings, needs, and preferences (Deng et al., 2022). As shown in Figure 1, the question asks the performance of the *Nikon binocular* in a low-light environment. A simple answer about the fact "*wide angle*" or one-sided opinion "*I think it's alright*" is hard to satisfy the various information needs of users. The users expect a good answer that can not only include relevant facts from the specifications of binoculars, but also cover both positive, neutral, and negative opinions from multiple perspectives. That provides them with a full understanding of product details and viewpoints. Answering this subjective question in an inductive way can be formalized as a challenging task (Pecar, 2018), which is called subjective induction QA, i.e., *SubjPQA*. Different from the factoid question, the answer in this task is more comprehensive, with multiple facts and diversified viewpoints from multiple data sources. It is hard for traditional extractive methods to integrate them.

In addition to multi-source heterogeneous summarization, this task has difficulties in data scarcity and domain imbalance. In real-world applications, some domains have rich labeled resources while others have few. For example, in the *SupQA* dataset (Zhang et al., 2023), the categories *electronic*, *home*, *sports* account for over 60% of the data, while the domains of *beauty*, *clothing* account for only 0.5%. It is labor-intensive to acquire data in all domains and annotate them. Considering existing methods are data-driven, insufficient labeled data would lead to under-training (Li et al.,

* Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9074–9089
August 11-16, 2024

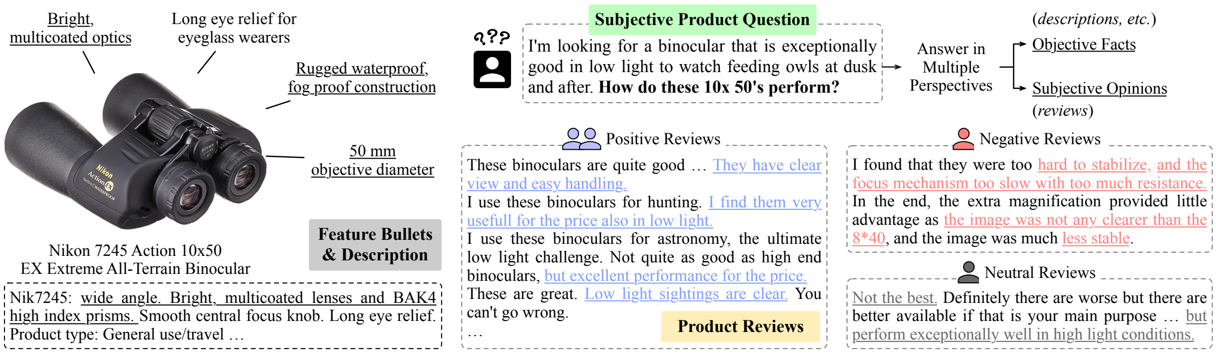

Figure 1: A subjective induction question example. The satisfactory answer should be able to aggregate relevant facts and users' diversified opinions of the product. Underlined texts are the key clues to answer the questions.

2022). That makes it hard to achieve reliable performance in the low-resource domains. A simple solution is knowledge transfer (Yu et al., 2021), which leverages the supervision from a rich source to supplement the poor one (Zou et al., 2021) by either pretrain-finetuning or multitask learning. However, both transfer methods require a large size of labeled data in the target domain (Zhao et al., 2022), but in reality, there is often very little target data. Although we can do pre-training with the help of large language models, it requires careful design of prompts for each domain, which is labor-intensive. Moreover, the context patterns and feature distributions are different between various domains. For example, '*screen*' and '*battery*' are common in the electronics domain, whereas '*fabric*' and '*fit*' are often used in the clothing domain. This domain shift problem would lead to a performance drop across different domains (Gu et al., 2021), especially when generating our domain-related answers that involve the induction of multi-viewpoints.

To address these challenges, we propose a new self-adaptive model for the *SubjPQA* task. The motivation is that we can disentangle some domain-agnostic knowledge from a large amount of source data to enhance the target generation under certain conditions. In detail, for each sample, we first acquire all answer-related implicit knowledge and encode their textual features. Two encoders are used for the source and target domains, respectively. To support knowledge transfer, the parameters of the source encoder are used to initialize the target one. We then project each representation into a latent space to disentangle two key factors. One is the invariant factor, which captures some inductive patterns and semantic expressions common to most domains; the other is the domain-specific factor to grasp the aspects and attributes unique to a certain domain. Such invariant representations enable us to flexibly reduce the domain impact when transferring from a rich source to a low-resource target. To support domain adaptation, we design three criteria for disentangled learning. The first one is based on reconstruction, where the disentangled representations should maximally preserve the information integrity. The second one is via adversarial learning. That is, a discriminator is used to distinguish the representations from different domains, while the encoders try to fool it inversely. Moreover, we use another criterion of contrastive learning to ensure that the reconstructed representations are valid. By multi-criteria joint learning, we can reduce the domain deviation to learn robust representations that can generalize well to long-tail domains. Finally, we use the learned latent vectors from the target domain to generate multi-perspective summaries as the answers for the *SubjPQA* task. Extensive experimental results from popular datasets demonstrate the effectiveness of our approach.

The main contributions of this paper include,

- We reveal the issue of imbalanced domain resources in the field of *SubjPQA*, and point out the challenges of adapting to low-resource domains, which are new for this task.
- We propose a new adaptive model based on adversarial disentangled learning that derives key latent factors to grasp domain generalization knowledge for low-resource *SubjPQA*.
- We conduct extensive experiments on the popular dataset to evaluate the rationality and effectiveness of our approach.

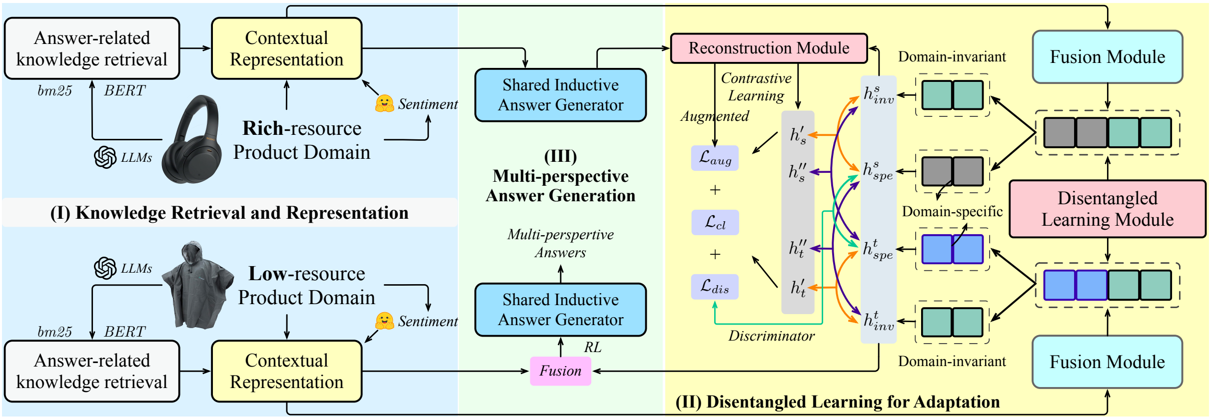

Figure 2: An overview architecture of the self-adaptive model for low-resource *SubjPQA*.

2 Methodology

We first give some notations for this task. Let $D_s = \{c_1, \ldots, c_{l_s}\}$ denote the contexts of products, with l_s utterances in the source domain, where $c_i = \{x_{i,1}, \ldots, x_{i,l_{s_i}}\}$ is the i-th context with l_{s_i} words. We use the special symbol $\langle sep \rangle$ to separate contexts. The gold answer contains an objective part $Y_s^{obj} = \{y_1, \ldots, y_{l_o}\}$ and a subjective part $Y_s^{subj} = \{y_1, \ldots, y_{l_s}\}$. Based on the target contexts $D_t = \{c_1, \ldots, c_{l_t}\}$, our goal is to yield this inductive answer with correct Y_t^{obj} and Y_t^{subj} with low labeled resource. Next, we illustrate the adaptive model for low-resource *SubjPQA* in Figure 2, which includes knowledge retrieval and representations, disentangled learning, and target model adaption.

2.1 Knowledge Retrieval and Representations

To build a *SubjPQA* model, we first need to acquire and represent the answer-related knowledge finely.

Subjective questions often involve various kinds of knowledge, such as product aspects, attributes, facts, opinions, and even implicit commonsense relations. For example, it is common sense that *refresh rate* is related to the *phone screen*. When users ask about the *refresh rate*, they actually want to know the quality of *phone screen*. Without capturing such knowledge, it is difficult to provide satisfactory answers. Thereby, we propose to use prompts to inquire about the external large language model (Nie et al., 2023), which is reported to contain rich implicit knowledge. This content can supplement the necessary but missing contexts of the inputs, enabling models to better derive the answer. Considering the content may contain irrelevant noise, we filter them by keyword matching (via the *BM25* (Askari et al., 2023)) and semantic similarity (via *BERT* (Zhang et al., 2019) model). Concretely, we use the *BM25* score f_{bm25} and *BERT* embedding-based f_{bert} to compute cosine similarity between product questions with other product contents. *BM25* has high precision but low recall, while *BERT* has higher recall. By combining them, we can learn from each other to accurately acquire the relevant content with high recall. The final score is a weighted sum of two scores, as Eq.(1),

$$f_{score} = \alpha \cdot \sigma(f_{bm25}) + (1 - \alpha) \cdot f_{bert}, \quad (1)$$

where α is a hyperparameter and σ is the sigmoid function to transforms the f_{bm25} into a value between 0 and 1. The content with a score f_{score} below a threshold ε is viewed as noise and filtered.

Based on the retrieved contents, we then derive contextual representations for the samples. Each sample consists of product descriptions, a set of reviews, and implicit knowledge. We then encode each textual part by *BART* (Lewis et al., 2020) which is effective in capturing fine-grained context. We input the question and each textual part to compute their cross-attention (Roy and Kundu, 2023) which can emphasize the question-aware content. Next, we grasp correlations among the textual parts by an inter-attention mechanism (Li et al., 2023c). We incorporate the dependent context to update token representations. This is done by concatenating all context vectors and applying a linear transformation with a trainable matrix W_c, as Eq.(2),

$$h_\tau^k = W_c \cdot [\overset{h}{\underset{i=1}{\|}} \sum\nolimits_s A_i(c_s, c_k)], \quad (2)$$

where $\|$ is the concatenating operation, $A_i(\cdot)$ is the attention function, h_τ^k is the k-th inter-attention vector. Finally, we fuse the vectors with cross-and-inter attention to obtain a representation vector h_τ.

2.2 Disentangled Learning for Adaptation

Different product domains have deviations in terms of feature patterns and label distributions, leading to the domain shift problem. In other words, the model trained in a source domain may be biased towards some subordinate features instead of key discriminant factors. The answer decoder will easily confuse to swing around two domains. That would deteriorate the performance and generalization ability of the model (Zhao et al., 2022), making it difficult to transfer to the low-resource domain (Chronopoulou et al., 2022). To handle this issue, we introduce a disentangled learning module that can separate latent factors, including the invariant factor shared across domains, and the one customized to a specific domain. In this latent space, we can narrow the gap between two domains and grasp the key cross-domain knowledge. That can improve the adaptability of the model and tackle the target data sparsity accordingly.

(1) Disentanglement. We observe the vector h_τ obtained from the previous step may be entangled with various biased correlations. To reduce the bias and enhance the model's discriminability, we disentangle it into two independent latent factors, i.e., invariant h^τ_{inv} and domain-specific h^τ_{spe}, as Eq.(3),

$$h^\tau_{inv} = \mathcal{M}_{inv}(h_\tau), \quad h^\tau_{spe} = \mathcal{M}_{spe}(h_\tau), \qquad (3)$$

where $\mathcal{M}_{inv}$ and $\mathcal{M}_{spe}$ are multi-layer feedforward networks with a *LeakyReLU* activation function.

(2) Reconstruction. To guide the disentangled direction, we regularize the distance of the two invariant vectors from the source and target domains, and let them be semantically closer. Simultaneously, we encourage their domain-specific vectors to be close in the same subspace. That allows the model to better transfer knowledge to the low-resource domain. In detail, we employ backtracked (Li et al., 2023a) and cross-domain (Yu et al., 2023a) reconstruction strategies. The first one is to restore the original representation h_τ based on the disentangled vectors h^τ_{inv} and h^τ_{spe}, as Eq.(4). The motivation is that a good disentanglement should maximally preserve information of the original vector.

$$h'_\tau = \mathcal{M}^{(1)}_{rec}([h^\tau_{inv}; h^\tau_{spe}]),\ \tau \in \{s, t\}, \qquad (4)$$

where τ denotes the source/target domain, h^τ_{inv} and h^τ_{spe} are the invariant and specific vectors, and h'_τ is the reconstructed vector. Further, we use the source invariant vector h^s_{inv} and a domain-specific vector h^t_{spe} from the target to reconstruct the source original vector h''_s, as Eq.(5). That can reduce the discrepancy caused by different domain data, thus enabling new domain adaption.

$$h''_s = \mathcal{M}^{(2)}_{rec}([h^s_{inv}; h^t_{spe}]). \qquad (5)$$

To facilitate joint learning of disentanglement and reconstruction, we adopt a data augmentation technique, which can provide data to start the iterative process. Satisfactory disentangled vectors should be able to decode the answers effectively. Thus, we utilize the reconstructed vectors h'_s, h''_s, h'_t and h''_t to yield answers $\hat{Y}^{obj}_s$, $\hat{Y}^{obj}_t$ in both the source and target domains. The reconstructed representations act as enhanced training signals to regularize the disentanglement. We define the augmented generation loss as the cross-entropy between the generated answers and the ground truth:

$$\begin{aligned} \mathcal{L}^{s,\omega_s}_{aug} &= -\textstyle\sum_{t=1}^{l^s_o} \log P(y_t|\hat{y}_1, \ldots, \hat{y}^*_{t-1}, \omega_s), \\ \mathcal{L}^{t,\omega_t}_{aug} &= -\textstyle\sum_{t=1}^{l^t_o} \log P(y_t|\hat{y}_1, \ldots, \hat{y}^*_{t-1}, \omega_t), \end{aligned} \qquad (6)$$

where $\omega_s \in \{h'_s, h''_s\}$ and $\omega_t \in \{h'_t, h''_t\}$ indicate the representations under different reconstruction strategies, and $l^{s/t}_o$ is the length of the gold objective answer. The augmented loss $\mathcal{L}_{aug}$ acts as an auxiliary training objective. That encourages the model to better learn transferable knowledge, so as to produce high-quality answers in both domains.

(3) Discriminator. To learn the optional disentangled mapping function, we employ adversarial learning by using a discriminator to distinguish between the domain-specific representations (h^s_{spe} and h^t_{spe}) from the source and target domains. This helps to reduce the distance in a shared subspace between representations of two domains, thus encouraging the reconstructed target distributions to be closer to the source one. In this way, the external domain data can be used to enhance the training of in-domain data. The discriminator is also implemented as a 3-layer feed-forward network. The logistic loss function is defined as Eq.(7).

$$\mathcal{L}_{dis} = -\sum\nolimits_{\tau \in \{s,t\}} \log P_{\mathcal{D}}(\tau | h^\tau_{spe}), \qquad (7)$$

where $P_{\mathcal{D}}(\tau|h^\tau_{spe})$ is the predicted probability of $\mathcal{D}$ that h^τ_{spe} belongs to the source or target domains.

(4) Contrastive Learning. To learn effective representation, we further employ contrastive learning (Gao et al., 2021) which is good at pulling semantically close neighbors together and pushing non-neighbors apart. The reconstructed vector is expected to have similar semantics with the original representation (Chen et al., 2020). We thus take the original and reconstructed representations from the target domain and treat them as a positive pair. Then we contrast the positive pairs against negative samples from the training batch. That can push away dissimilar representations, which helps the model better learn domain-invariant knowledge. The contrastive loss is defined as Eq.(8).

$$\mathcal{L}_{cl}^{(1)} = -\sum_{i=1}^{n} \log \frac{e^{\mathrm{sim}(h_t^{(i)}, h_t'^{(i)})/\gamma}}{\sum_k \mathbb{1}_{[k\neq i]} e^{\mathrm{sim}(h_t^{(i)}, h_t'^{(k)})/\gamma}},$$
$$\mathcal{L}_{cl}^{(2)} = -\sum_{i=1}^{n} \log \frac{e^{\mathrm{sim}(h_t^{(i)}, h_t''^{(i)})/\gamma}}{\sum_k \mathbb{1}_{[k\neq i]} e^{\mathrm{sim}(h_t^{(i)}, h_t''^{(k)})/\gamma}}, \tag{8}$$

where $\mathbb{1}$ is the indicator function, $\mathrm{sim}(\cdot)$ is a cosine similarity function, n is the number of batch size, and γ is the temperature parameter. In addition, the reconstruction vectors (h_s', h_s'') in the source domain are expected to be similar, but dissimilar to all other instances in the same training batch. Thus, we have the loss as Eq.(9),

$$\mathcal{L}_{cl}^{(3)} = -\sum_{i=1}^{n} \log \frac{e^{\mathrm{sim}(h_s'^{(i)}, h_s''^{(i)})/\gamma}}{\sum_k \mathbb{1}_{[k\neq i]} e^{\mathrm{sim}(h_s'^{(i)}, h_s''^{(k)})/\gamma}}. \tag{9}$$

Based on the two contrastive learning strategies in Eq.(8) and Eq.(9), we can obtain more robust disentangled representations that effectively share common knowledge across domains while retaining domain-specificity. This allows better adaptation and performance for *SubjPQA* in the low-resource domains. Finally, the contrastive loss over the training batch can be formulated as Eq.(10),

$$\mathcal{L}_{cl} = \mathcal{L}_{cl}^{(1)} + \mathcal{L}_{cl}^{(2)} + \mathcal{L}_{cl}^{(3)}. \tag{10}$$

By combining adversarial learning and contrastive learning, the joint loss function for the domain adaptation part is defined as Eq.(11), where β_1 and β_2 are the hyper-parameters to adjust the joint loss.

$$\mathcal{L}' = \beta_1 \mathcal{L}_{dis} + \beta_2 \mathcal{L}_{cl} + (1-\beta_1-\beta_2)\mathcal{L}_{aug},$$
$$\mathcal{L}_{aug} = \sum\nolimits_{\tau\in\{s,t\}} (\mathcal{L}_{aug}^{\tau,\omega_s} + \mathcal{L}_{aug}^{\tau,\omega_t}). \tag{11}$$

2.3 Multi-perspective Answer Generation

For the *SubjPQA* task, a good answer that can meet users' information needs should include two parts, including objective product facts and subjective opinions based on user reviews. We design a decoder to yield this multi-perspective inductive result. In particular, we first concatenate four kinds of vectors to facilitate knowledge transfer, including the source invariant vector h_{inv}^s, the target invariant vector h_{inv}^t, domain-specific vector h_{spe}^t and contextual vector h_t from the source and target domain, respectively. By feeding them into a linear transformation, we can obtain $\boldsymbol{h}_f$. Based on it as input along with the previous words, the decoder produces the answer word-by-word. The decoder is learned separately through two parts: the objective part is trained via maximum likelihood estimation (*MLE*) on product facts. For the subjective part, we design a template according to the reviews' sentiment distribution to cover diverse viewpoints with positive, negative, and neutral opinions. To train the whole model, we employ reinforcement learning (Yadav et al., 2021), with the following process:

• *Objective Part.* To generate the objective part of the inductive answers, we utilize the disentangled vectors (h_{inv}^t and h_{spe}^t) and contextual vectors h_t to input the decoder. The model is trained based on the negative log-likelihood loss, as Eq.(12).

$$\mathcal{L}_{mle} = -\sum_{t=1}^{l} \log(w_t^* | w_1^*, \ldots, w_{t-1}^*, \boldsymbol{h}_f), \tag{12}$$

where w_j^* denotes the decoder output at the step j.

• *Subjective Part.* We feed the fused representation $\boldsymbol{h}_f$ into the decoder. Besides, we design two reward functions to encourage the decoder to learn the expected answering way. One is the sentiment recognition reward r_{sen} which ensures the generated answers reflect diverse viewpoints with various sentiments. Another is the template-identified reward r_{tem} which is introduced to help the verbalism of results comply with the pre-defined expressive structure in the template. The model is trained via policy gradient methods to maximize a mixed reward. We minimize the RL loss function as:

$$\mathcal{L}_{rl} = -\mathbb{E}_{\mathcal{A}^s\sim p_\theta}[r(\mathcal{A}^s, \mathcal{A}^*, \mathcal{T}^*)], \tag{13}$$

where $\mathcal{A}^*$ is the gold answer, $\mathcal{A}^s$ is the answer by sampling the words from the model's output distribution, and $\mathcal{T}^*$ is the template. The overall reward $r(\mathcal{A}^s, \mathcal{A}^*, \mathcal{T}^*)$ is the weighted sum of sentiment recognition and template identification rewards. Finally, we adopt policy gradient methods to train the RL-based answer decoder. The network is trained using the mixed loss as Eq.(14).

$$\mathcal{L}'' = \eta\mathcal{L}_{rl} + (1-\eta)\mathcal{L}_{mle}, \tag{14}$$

where η is the scaling factor, which is used to balance the weights between the reward loss and the maximum likelihood estimation loss. In this way, our model can produce comprehensive summaries with multi-perspective viewpoints as inductive answers for the low-resource domain.

3 Evaluations

We extensively evaluated the effectiveness of our method with quantitative and qualitative analysis.

3.1 Data and Experimental Settings

To evaluate our proposed method, we utilized a typical *SupQA* dataset (Zhang et al., 2023) in the field of *SubjPQA* which contained $48,352$ samples across 15 product domains. Each sample included a subjective question, product descriptions, attributes, multiple user reviews with diversified sentiments, and a multi-perspective answer summarizing objective facts and subjective reviews. The dataset covered a diverse range of product categories with imbalanced distributions, enabling us to evaluate domain adaptation capabilities. This made it more suitable than other datasets to evaluate our task. For example, other *PQA* datasets (i.e., *Amazon* (Wan and McAuley, 2016), *SubjQA* (Bjerva et al., 2020), and *AmazonQA* (Gupta et al., 2019)) could only provide simple answers in one-side, which did not reflect the multi-viewpoint summarization to suit our task. Thus, the *SupQA* dataset was representative and can effectively evaluate the model. More details were given in Appendix A. We further used *BLEU* (B_n) (Papineni et al., 2002) and *ROUGE* (R_n) (Lin, 2004) to measure the generated quality of the inductive answer. These metrics were widely used in the field of text generation, where our answer is essentially a summary of multiple contents. We repeated running 10 times and reported the average performance to reduce bias. The metric values would be larger when the generated content can well resemble a human speaking way.

To simulate the low-resource scenarios, we selected minimum amounts, medium amounts, and maximum amounts (that is, 1%, 5%, and 10%, respectively) of training samples randomly. The samples from the *Sports and Outdoors* (*SO*) and *Video Games* (*VG*) were treated as target domains, while the full *Electronics* (*ET*) and *Home and Kitchen* (*HK*) as sources. That created an imbalanced distribution across domains, with sparse *SO* and *VG*.

3.2 Comparisons Against State-of-the-Arts

We compared our method against 8 mainstream models, including (1) *BM25* (Robertson et al., 2009) and *LexRank* (Erkan and Radev, 2004), which were typical retrieval-based methods and acquired highly relevant product-related contents for each question; (2) *PGNet* (See et al., 2017), a robust *Seq2seq* model that utilized a hybrid pointer-generator mechanism to copy words directly from the source text; (3) *BART* (Lewis et al., 2020), a strong denoising autoencoder for many generation tasks; (4) *Pegasus* (Zhang et al., 2020a), a powerful transformer-based model with good performance on both rich and low-resource summarization tasks; (5) *InstructDS* (Wang et al., 2023), a query-based summarization model with instruction-following capabilities to generate tailored summaries; (6) *PlanSum* (Amplayo et al., 2021) and *FewSum* (Bražinskas et al., 2020), two typical opinion summarization models, which generated the subjective answers by summarizing product reviews. We reimplemented these baselines following their original settings, as shown in Appendix B.

The inductive answer covers both the objective facts and subjective opinions of the product. As illustrated in Table 1, for the objective part of the answer, our model was superior to the benchmark method (i.e., *BART*) under various low-resource settings. Based on only 1% target data, our model obtained improvements of + 5.47 and + 6.50 in terms of *BLEU-2* and *ROUGE-L*, respectively, as compared to *InstructDS*. When we increased the training data, our advantage was still significant. This indicated our learned disentangled representations could effectively transfer knowledge from the rich domain to the low-resource one. In addition, for the subjective part of the answer, our model still significantly outperformed the best baseline (i.e., *InstructDS*). That showed the effectiveness of our model on learning subjective knowledge. We observed finetuned models (i.e., *PGNet*, *BART* and *Pegasus*) fell short in comparison to the instruct-tuning model *InstructDS*. More results of other product domains are detailed in Appendix C.1.

3.3 Human Evaluations

Moreover, we conducted human evaluations to qualitatively evaluate the answer's quality. To finely analyze the inductive answer, we employed four metrics, including *Factness* (*Fact*) and *Accuracy* (*Acc*) to measure its objective part, *Comprehensiveness* (*Comp*) and *Template Compliance* (*Tcp*) for subjective part. *Fact* evaluated the coverage of the answer-related facts, while *Acc* measured the accuracy of to-the-point facts in the answer.

Table 1: Comparisons of all evaluated methods on the different proportions of training data. ET is the source domain while SO is the low-resource target domain. The results were significant using a statistic t-test with p-value<0.005.

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$
BM25	11.77	26.91	11.73	26.88	11.75	26.84	2.97	12.67	2.91	12.63	2.95	12.65
LexRank	11.93	27.09	11.89	27.06	11.91	27.02	3.03	12.82	2.97	12.78	3.01	12.80
PGNet	3.52	8.78	3.55	8.91	3.79	9.08	3.58	8.12	3.80	8.33	4.01	8.64
BART	22.01	28.12	22.05	29.41	23.91	30.67	25.34	29.31	26.77	30.63	25.89	30.01
Pegasus	25.43	31.41	26.53	32.48	27.31	34.01	29.63	34.02	29.97	34.87	30.41	35.30
InstructDS	27.89	33.49	28.82	34.71	29.64	36.13	47.19	50.48	48.34	51.61	49.40	52.73
PlanSum	-	-	-	-	-	-	4.88	12.70	4.97	12.81	5.01	12.87
FewSum	-	-	-	-	-	-	6.81	15.08	6.93	15.21	7.06	15.34
Ours	**32.36**	**38.99**	**33.19**	**39.65**	**34.21**	**40.89**	**53.11**	**57.17**	**54.39**	**58.23**	**55.54**	**59.28**

Table 2: Human evaluation results on different low-resource scenarios (source domain: ET; target domain: SO). Statistically significant with t-test, p-value<0.005. The performance of the other domains is shown in Appendix C.3.

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑
Pegasus	0.57	0.59	0.61	0.62	0.64	0.63	0.50	0.71	0.52	0.71	0.54	0.72
InstructDS	0.63	0.64	0.64	0.65	0.66	0.67	0.64	0.85	0.65	0.86	0.66	0.87
Ours	**0.72**	**0.71**	**0.74**	**0.74**	**0.77**	**0.78**	**0.76**	**0.90**	**0.78**	**0.92**	**0.80**	**0.92**

Comp checked whether the answer could fully summarize various viewpoints, and *Tcp* assessed the fluency of the answer and its matching degree to the pre-defined template. A 3-point range was used for each metric, with $[0, 0.33]$ being low, $[0.34, 0.66]$ being medium, and $[0.67, 1]$ being high. To avoid biases, we randomly sampled test cases to grade the judges by participants. We employed *Randolph's kappa* to measure the inter-rater reliability. The *kappa* scores κ were all higher than 0.7, which indicated a good agreement. As shown in Table 2, our model significantly outperformed other baselines in terms of all metrics. That was consistent with the quantitative results in the previous section. These results reflected that our model could disentangle key latent factors to better transfer objective and subjective knowledge across product domains. More details were provided in Appendix C.2.

3.4 Ablation Studies

To analyze the contribution of each component in our model, we conducted ablation studies by removing four key modules from our framework one by one, including (1) *DisM* that dropped the disentangled learning module and solely relied on *BART* encoders; (2) *ConL* discarded the contrastive loss objective in Eq.(9); (3) *RecM* that removed back-tracked and cross-domain reconstruction strategies; (4) *DomD* that threw away a domain discriminator.

As shown in Figure 3, removing the disentangled module caused the most significant performance drop, indicating it played a crucial role in transferring knowledge across domains. Besides, excluding other three modules also led to noticeable degradation. These results verified all components in our model were beneficial for obtaining key factors to build a robust model. More evaluations on the disentangled module and case study were illustrated in Appendix 3.5, 3.6, respectively.

3.5 Study of Disentangled Representation

Besides, we utilized the t-SNE (Van der Maaten and Hinton, 2008) algorithm to visualize the latent representations after the disentangled learning and reconstruction. We used *ET* and *HK* as rich-resource domains and *SO* and *VG* as low-resource domains. As presented in Figure 4, after disentanglement, the demarcation between the two domains was clearer. This separation enabled the model to finely capture key discriminant factors, thus enhancing the decoding ability. After reconstruction with adversarial and contrastive learning, the hidden subspaces gradually converged and became more aligned across domains. This validated the ef-

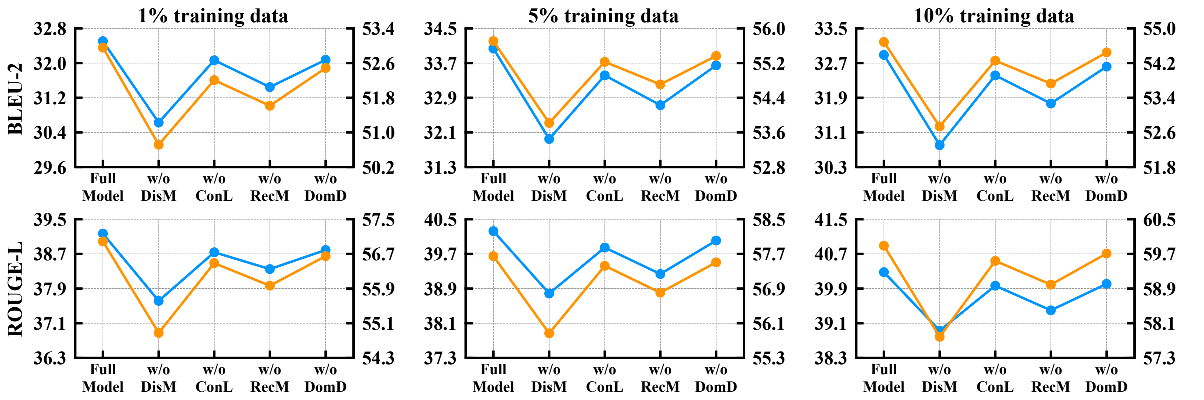

Figure 3: Ablation studies with different low-resource settings (source domain: ET; target domain: SO). Orange and Blue lines represent the objective and subjective summarized parts of the inductive answers, respectively.

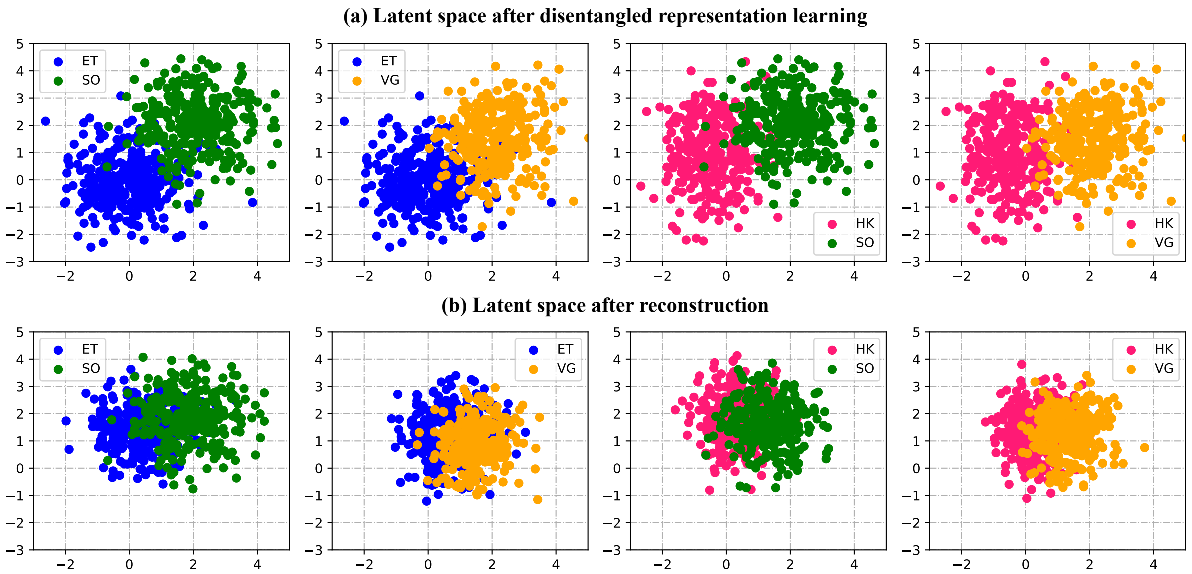

Figure 4: Visualization of the disentangled representations across domains. Subgraph (a) shows the representations after the disentangled learning. Subgraph (b) displays the representations after reconstruction.

fectiveness of extracting invariant knowledge while retaining domain-specificity for domain transfer.

3.6 Case Studies and Discussions

To provide more insights into our model, we conducted case studies, including one from the *ET* source domain and another from a low-resource *Sports and Outdoors* (*SO*) domain. As shown in Table 5, our model can generate satisfactory inductive answers for subjective questions in the low-resource domain. With only 5% training data from the target *SO* domain, our model still can produce an inductive answer capturing the key product facts like weight limit and dimensions. That indicated effective transfer of factual knowledge from the source *ET* via the learned domain-invariant representations. For the subjective part of the answer, our model roughly followed the template structure to aggregate diversified users' opinions. The generated answers covered positive, neutral, and negative viewpoints on the suitability of the *kayak*. This showed our model could produce comprehensive answers aggregating multiple viewpoints.

4 Related Work

Product Question Answering (*PQA*) has received extensive attention in recent years (Deng et al., 2023). Earlier efforts try to answer the yes/no questions about users' opinions (McAuley and Yang, 2016), such as "*Is the user satisfied with the mobile phone?*" That can be framed as opinion mining (Yu and Lam, 2018), and answered by classifying the opinion polarity. On the other hand, some researchers focused on the factoid questions, where

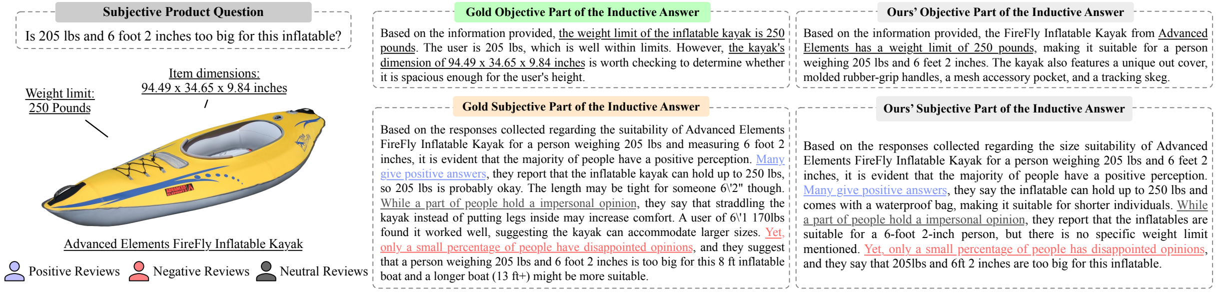

Figure 5: A case study with only 5% of target data from the '*Sports and Outdoors*' target domain. Different color highlights represent different sentiment perspectives. Black underlines indicate key objective factual information.

their answers were often text spans in the given document (Xu et al., 2019). To extract the answers, many methods based on machine reading comprehension (Cui et al., 2017) were proposed. Another method was based on retrieval, which treated *PQA* as an answer selection task (Yu et al., 2018). The documents were ranked from a set of candidates to select the most relevant one as results (Zhang et al., 2020b). In addition, some works proposed to generate the answers directly based on the *Seq2seq* framework (Gao et al., 2019). To answer the opinion questions, Deng et al. (2020) proposed to produce opinion-aware answers based on multi-task learning. To capture the fine-grained relations of opinions and facts, Feng et al. (2021) utilized a heterogeneous graph neural network to form subjective answers. Besides, there are other works on *PQA*, such as analyzing the semi-structured data (Shen et al., 2022), making cross-lingual prediction (Shen et al., 2023). Due to the lack of consideration of the answers' structure, existing results often only cover either a piece of facts or opinions, which are not comprehensive.

Low-resource Generation(Yu et al., 2020) has gained significant attention in the literature (Yu et al., 2023c). Yu et al. (2021) was the first to study the domain adaptation (Li et al., 2023b) for generation. Chen and Shuai (2021) introduced an effective meta-transfer learning method for low-resource summarization. Besides, Cheng et al. (2023) proposed a reinforcement method to support generation in new domains with limited data. Sukhadia and Umesh (2023) designed a domain adaptive method, which used encodings of a pre-trained *ASR* model as features to learn a target model. Calderon et al. (2022) presented a corrupt-and-reconstruct approach for generating domain counterfactuals and applied it as a data augmentation method. Zhao et al. (2022) explored a prompt learning model to handle zero-resource summarization. Zhang et al. (2022) developed a *GAN*-based framework for one-shot domain adaptation, leveraging a reference image and its binary entity mask to transfer pre-trained *GAN* styles and entities to a target domain with minimal data(Yu et al., 2023b). In contrast, we propose a new adaptive model that derives key latent factors with several constraints to grasp domain generalization knowledge, which helps to achieve low-resource adaption.

5 Conclusion

This paper studied the task of subjective question answering on products (*SubjPQA*). We revealed the issue of domain bias and imbalance, where the patterns and data distributions would vary in different domains. That posed challenges for traditional methods to achieve reliable performance on low-resource domains. To tackle this problem, we proposed to transfer knowledge from a rich domain to a poor one, and designed a novel self-adaptive model to facilitate transferring based on adversarial disentangled learning. In particular, for each instance in the source and target domain, we first retrieved answer-related knowledge and represented their contexts. We then disentangled their representations into domain-invariant and domain-specific latent vectors. To guide the disentangled direction, we designed the backtracked and cross-domain reconstruction constraints, which can regularize the results to maximally preserve the original data characteristics and capture their cross-domain correlations. Moreover, we developed an adversarial discriminator with contrastive learning to reduce the impact of out-of-domain bias. Based on learned latent vectors for the target domain, we decoded summaries multi-perspective viewpoints as inductive answers for *SubjPQA*. Extensive experiments on the typical dataset demonstrated the effectiveness of our method in low-resource settings.

6 Acknowledgments

This work is supported by the National Natural Science Foundation of China (62276279, 62102463, 62372483, 62276280, U2001211, U22B2060), Guangdong Basic and Applied Basic Research Foundation (2024B1515020032), Research Foundation of Science and Technology Plan Project of Guangzhou City (2023B01J0001, 2024B01W0004), Tencent WeChat Rhino-Bird Focused Research Program (WXG-FR-2023-06), and Technology Innovation Center for Collaborative Applications of Natural Resources Data in GBA, MNR (2024NRDZ01).

Limitations

The subjective induction QA task in a low-resource scenario is challenging. Although we had proposed an effective model, there is still room for improvement. Current model is few-shot but not zero-shot. Zero-shot learning for domain transfer remains an open challenge. That is a promising research direction to realize domain alignment without using any target labeled data. In addition, by analyzing our bad cases, the results still had some mistakes, such as temporal errors, typos, adverb errors, etc. These challenges will be studied in future works.

Ethics Statement

This paper aims to generate an inductive answer for the subjective question on products. Excluding the misusage scenarios, there are few or even no ethical issues with this technology. However, the framework is based on product reviews. It is possible to input some low-quality reviews related to the moral issue topics, resulting in some offensive results. We have taken into account this matter. With diverse viewpoints in the input reviews, we summarize their sentiment as positive, negative, and neutral. For each sentiment category, our model summarizes key aspects and attributes to yield answers. These answers contain knowledge from multiple perspectives, including relevant facts, overall sentiment, positive representative viewpoints, negative ones, and neutral ones, etc. In this classify-then-summarize way, we can aggregate heterogeneous and inconsistent viewpoints as a comprehensive answer. We observe that low-quality reviews are usually uninformative, coarse, and unimportant, while high-quality reviews are often informative and full of fine-grained details. Our summarizer is designed based on maximum salient information coverage. That can help to aggregate the salient aspects of high-quality reviews, and alleviate the effect of low-quality content.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. In *Proceedings of the AAAI conference on artificial intelligence*, 14, pages 12489–12497, Vancouver, Canada.

Arian Askari, Amin Abolghasemi, Gabriella Pasi, Wessel Kraaij, and Suzan Verberne. 2023. Injecting the bm25 score as text improves bert-based re-rankers. In *Proceedings of the 45th European Conference on Information Retrieval, ECIR*, pages 66–83, Dublin, Ireland.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4119–4135, Online.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746, Dublin, Ireland.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1597–1607, Vienna, Austria.

Yi-Syuan Chen and Hong-Han Shuai. 2021. Meta-transfer learning for low-resource abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14, pages 12692–12700, Online.

Pengsen Cheng, Jinqiao Dai, Jiamiao Liu, Jiayong Liu, and Peng Jia. 2023. Reinforcement learning for few-shot text generation adaptation. *In Journal of Neurocomputing*, 558:126689.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1351, Seattle, United States.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A customer service chatbot for E-commerce websites. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 97–102, Vancouver, Canada.

Yang Deng, Yaliang Li, Wenxuan Zhang, Bolin Ding, and Wai Lam. 2022. Toward personalized answer generation in e-commerce via multi-perspective preference modeling. *In Journal of ACM Transactions on Information Systems (TOIS)*, 40(4):1–28.

Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Opinion-aware answer generation for review-driven question answering in e-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 255–264, Virtual Event, Ireland.

Yang Deng, Wenxuan Zhang, Qian Yu, and Wai Lam. 2023. Product question answering in E-commerce: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11951–11964, Toronto, Canada.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *In Journal of artificial intelligence research*, 22:457–479.

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

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond i.i.d.: Three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, page 3477–3488, Ljubljana, Slovenia.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4996–5002, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Warut Khern-am nuai, Hossein Ghasemkhani, Dandan Qiao, and Karthik Kannan. 2023. The impact of online q&as on product sales: The case of amazon answer. *In Journal of Information Systems Research*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

Jinpeng Li, Yingce Xia, Xin Cheng, Dongyan Zhao, and Rui Yan. 2023a. Learning disentangled representation via domain adaptation for dialogue summarization. In *Proceedings of the ACM Web Conference 2023*, page 1693–1702, Austin, TX, USA.

Jinpeng Li, Yingce Xia, Xin Cheng, Dongyan Zhao, and Rui Yan. 2023b. Learning disentangled representation via domain adaptation for dialogue summarization. In *Proceedings of the ACM Web Conference*, pages 1693–1702, Austin, TX, USA.

Junjie Li, Jianfei Yu, and Rui Xia. 2022. Generative cross-domain data augmentation for aspect and opinion co-extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4219–4229, Seattle, United States.

Miao Li, Eduard Hovy, and Jey Han Lau. 2023c. Towards summarizing multiple documents with hierarchical relationships. *arXiv preprint arXiv:2305.01498*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635, Montreal, Canada.

Weizhi Nie, Yuru Bao, Yue Zhao, and Anan Liu. 2023. Long dialogue emotion detection based on commonsense knowledge graph guidance. *In Journal of IEEE Transactions on Multimedia*, pages 1–15.

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

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA.

Samuel Pecar. 2018. Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, pages 1–8, Melbourne, Australia.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *In Journal of Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Prasenjeet Roy and Suman Kundu. 2023. Review on query-focused multi-document summarization (qmds) with comparative analysis. *In Journal of ACM Computing Surveys*, 56(1):1–38.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083, Vancouver, Canada.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the International Conference on Machine Learning*, pages 4596–4604, Stockholm, Sweden.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adria De Gispert. 2023. xPQA: Cross-lingual product question answering in 12 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 103–115, Toronto, Canada.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, and Adrià Gispert. 2022. semiPQA: A study on product question answering over semi-structured data. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 111–120, Dublin, Ireland.

Vrunda N. Sukhadia and S. Umesh. 2023. Domain adaptation of low-resource target-domain models using well-trained asr conformer models. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 295–301, Doha, Qatar.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *In Journal of machine learning research*, 9(11).

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

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2324–2335, Minneapolis, Minnesota.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 249–255, Online.

Jianfei Yu, Qiankun Zhao, and Rui Xia. 2023a. Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1456–1470, Toronto, Canada.

Jianxing Yu, Wei Liu, Shuang Qiu, Qinliang Su, Kai Wang, Xiaojun Quan, and Jian Yin. 2020. Low-resource generation of multi-hop reasoning questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6729–6739, Online.

Jianxing Yu, Qinliang Su, Xiaojun Quan, and Jian Yin. 2023b. Multi-hop reasoning question generation and its application. *In Journal of IEEE Trans. Knowl. Data Eng.*, 35(1):725–740.

Jianxing Yu, Shiqi Wang, Libin Zheng, Qinliang Su, Wei Liu, Baoquan Zhao, and Jian Yin. 2023c. Generating deep questions with commonsense reasoning ability from the text by disentangled adversarial inference. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 470–486, Toronto, Canada.

Qian Yu and Wai Lam. 2018. Review-aware answer prediction for product-related questions incorporating aspects. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 691–699, Marina Del Rey, CA, USA.

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

International Conference on Computational Linguistics, pages 2192–2203, Santa Fe, New Mexico, USA.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904, Online.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339, Online.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020b. Answer ranking for product-related questions via multiple semantic relations modeling. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 569–578, Virtual Event, China.

Yufeng Zhang, Meng-xiang Wang, and Jianxing Yu. 2023. Answering subjective induction questions on products by summarizing multi-sources multi-viewpoints knowledge. In *Proceedings of 2023 IEEE International Conference on Data Mining (ICDM)*, Shanghai, China.

Zicheng Zhang, Yinglu Liu, Congying Han, Tiande Guo, Ting Yao, and Tao Mei. 2022. Generalized one-shot domain adaptation of generative adversarial networks. In *Proceedings of the 35th Advances in Neural Information Processing Systems*, New Orleans, LA, USA.

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Ruotong Geng, Huixing Jiang, Wei Wu, and Weiran Xu. 2022. Adpl: Adversarial prompt-based domain adaptation for dialogue summarization with knowledge disentanglement. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 245–255, Madrid, Spain.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic.

A Statistic of the Dataset

There has existed a large-scale dataset *SupQA* tailored for *SubjPQA*. As shown in Figure 6, the distribution of product categories in the *SupQA* is imbalanced with categories *electronic*, *home*, *sports* accounting for over 60% of the data. Categories such as *beauty*, *clothing* account for only 0.5%. The shortage of labeled data makes it challenging for models to achieve reliable performance on *SubjPQA* in the low-resource product domains.

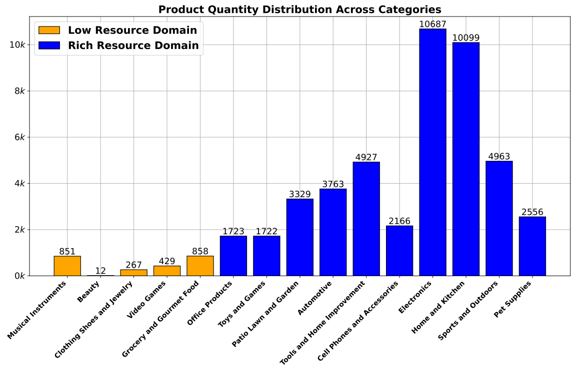

Figure 6: The distribution of product categories.

B Additional Implementation Details

For evaluation, we reimplemented each baseline with default settings. For fair comparisons, we conducted five runs and showed the average results.

Ours: We implemented our experiments on two Nvidia RTX 3090 GPUs with *PyTorch*. We initialized our model with the pre-trained $BART_{base}$ provided by *Hugging Face* (Wolf et al., 2020). The training batch size was 4, and we used the *Adam* optimizer with a learning rate of 3e-5 for $BART_{base}$ related modules and 1e-3 for the contextual representation module. During the adaptation step, we used a learning rate of 3e-6 for *MLE* training, 3e-7 for *RL* training, and 3e-4 for the discriminators. Our disentangled and reconstruction modules were built by multi-layer perceptron and activator. We utilized the grid search to tune the hyper-parameters according to the validation performance. The trade-off parameters α, ε, β_1, β_2, and η were set to 0.3, 0.9, 0.25, 0.5, and 0.9, respectively. The *LLM* we used to yield implicit commonsense knowledge was the *GPT*-3.5 turbo . The sentiments of product reviews in Figure 2 were obtained via a review text classifier on *Huggingface* platform, which was fine-tuned on Amazon product reviews corpus.

PGNet: For training the pointer-generator network model, we utilized a bidirectional two-layer *LSTM* as the encoder and a uni-directional single-layer *LSTM* as the decoder. The hidden size was

https://github.com/MetaZ1/ACL-SubjPQA

https://platform.openai.com/docs/guides/text-generation set to 512 to capture richer context representations. We initialized both models with 300-dimensional *GloVe* word embeddings pre-trained on a large corpus. To prevent overfitting, we set the maximum number of epochs to 30 and employed an early stopping strategy with a patience of 5 epochs. The gradient norm was clipped at 5.0 to stabilize training. The batch size was 64 for efficient training on GPUs, and we used the *Adam* optimizer with a learning rate of 2e-4 and a weight decay of 1e-5.

BART: We initialized the model with *base* size pre-trained on a large corpus for stronger language modeling capabilities. The maximum number of epochs was 20 with early stopping after 3 epochs if the validation loss does not decrease. The batch size was 32, and we utilized the *AdamW* optimizer with a linear warmup and decay schedule. The peak learning rate was 5e-5 and the weight decay was 1e-2 to regularize training objectives.

Pegasus: We reimplemented the model based on the *transformers* library with configurations for abstractive summarization tasks. The model featured a vocabulary size of 50,265 and utilizes 1024 position embeddings to handle long product-related content. It was composed of 12 layers each for the encoder and decoder, with a hidden size of 1024 and an *FFN* dimension of 4096. The model used 16 attention heads in both the encoder and decoder. We employed *GELU* as its activation function, with a dropout rate of 0.1 and an attention dropout of 0.0 to prevent overfitting. The model initialization standard deviation was set at 0.02, and it includes parameters for padding, end-of-sequence, and forced end-of-sequence tokens. For optimization, both pre-training and fine-tuning used *Adafactor* (Shazeer and Stern, 2018) with square root learning rate decay and dropout rate of 0.1.

InstructDS: The model was built on the foundation of *Flan-T5-XL* (Chung et al., 2022), leveraging *LORA* for parameter-efficient training, resulting in 37.7 million trainable parameters. This model was unique in its approach to dialogue summarization, aiming to synthesize high-quality query-based summarization triples by exploiting the question generation and answering capabilities of large language models. It underwent instruction tuning with a focus on general summarization, query-based summarization, and length-aware augmentations, making it adept at producing summaries that were tailored to user instructions and preferences.

For the **PlanSum** and **FewSum** subjective models, we followed the settings described in their original papers to ensure a fair comparison.

In the validation phase, we evaluated the loss for each epoch. When the loss was minimum, we can derived an optimal model for evaluation.

C Additional Evaluations

Due to the page limit, we showed additional experiments as follows, including the comparison results, human evaluation, disentangled module analysis, case study, and other impact aspects.

C.1 Extra Performance Comparisons

In this section, we show additional performance in various domains, including *ET* domain to *VG* domain, *HK* domain to *SO* domain, and *HK* domain to *VG* domain. As shown in Table 3, Table 4 and Table 5, we have the following observations:

• In terms of the evaluation metrics *BLEU-2* and *ROUGE-L*, our model significantly outperformed all baselines in different low-resource scenarios (using 1%, 5%, 10% target training data), for both objective and subjective part of the inductive answers. This demonstrated our model could effectively learn transferable knowledge from the source domain and apply it to the target domain.

• Even in the extremely low-resource case (only 1% target data), our model still obtained considerable improvements compared to other baselines. As available training data increased, the advantage of our model still remained quite significant. That further verified the efficacy of our disentangled representation learning.

C.2 Human Evaluation Settings

To ensure reliable and unbiased human evaluation, we recruited 8 undergraduate students majoring in computer science as annotators. All of them are good at English and have strong language skills. We conducted qualification tests on their annotation abilities before recruitment. Each annotator evaluated around 250 randomly sampled answers, with over 2,000 annotated answers in total. Comprehensive guidelines were provided detailing the scoring criteria, scales, and examples. The key metrics included *Factness*, *Accuracy*, *Comprehensiveness*, and *Template Compliance*. We calculated inter-annotator agreement using *Randolph's kappa*

https://github.com/rktamplayo/PlanSum
https://github.com/abrazinskas/FewSum

Table 3: Comparisons of all the evaluated methods on different low-resource scenarios. Here, we use the ET domain as the source domain and VG as the low-resource domain. Statistically significant with t-test, p-value<0.005.

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$
BM25	11.86	26.99	11.82	27.12	11.91	27.24	2.95	12.79	2.96	12.61	3.05	12.81
LexRank	12.02	27.21	12.05	27.31	12.11	27.41	3.15	12.92	3.12	12.86	3.19	12.97
PGNet	3.63	8.91	3.68	9.02	3.81	9.17	3.68	8.23	3.85	8.41	4.07	8.72
BART	22.31	28.34	22.42	29.63	24.02	30.87	25.51	29.47	27.02	30.78	26.13	30.24
Pegasus	25.67	31.64	26.81	32.73	27.53	34.21	29.84	34.27	30.21	35.02	30.67	35.51
InstructDS	28.13	34.68	29.02	35.91	30.01	36.41	47.41	50.72	48.51	51.87	49.67	53.02
PlanSum	-	-	-	-	-	-	4.97	12.81	5.12	12.93	5.17	13.01
FewSum	-	-	-	-	-	-	6.96	15.23	7.15	15.41	7.22	15.51
Ours	**32.55**	**39.17**	**33.41**	**39.87**	**34.37**	**41.03**	**53.36**	**57.41**	**54.61**	**58.37**	**55.81**	**59.41**

Table 4: Comparisons of all the evaluated methods on different low-resource scenarios. Here, we use the HK domain as the source domain and SO as the low-resource domain. Statistically significant with t-test, p-value<0.005.

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$
BM25	11.61	26.84	11.68	26.91	11.72	26.98	2.90	12.51	2.92	12.63	2.97	12.71
LexRank	11.72	26.91	11.79	27.01	11.83	27.05	2.95	12.61	2.98	12.69	3.01	12.77
PGNet	3.41	8.61	3.49	8.79	3.59	8.91	3.51	7.97	3.62	8.11	3.79	8.24
BART	21.92	27.97	22.11	29.21	23.69	30.61	25.12	29.21	26.49	30.41	25.93	29.87
Pegasus	25.21	31.22	26.31	32.49	27.09	33.87	29.51	33.88	29.92	34.71	30.32	35.11
InstructDS	27.83	33.34	28.71	34.61	29.51	35.97	46.99	50.41	47.92	51.72	48.79	52.87
PlanSum	-	-	-	-	-	-	4.82	12.51	4.89	12.62	4.97	12.69
FewSum	-	-	-	-	-	-	6.71	14.97	6.79	15.11	6.92	15.23
Ours	**31.99**	**38.71**	**32.79**	**39.53**	**33.68**	**40.76**	**52.81**	**56.98**	**53.92**	**58.11**	**54.93**	**59.21**

to ensure consistency. The agreement scores were 0.81 for *Factness*, 0.79 for *Accuracy*, 0.77 for *Comprehensiveness*, and 0.74 for *Template Compliance*, indicating reliable annotations. To avoid bias, all evaluated samples were randomly shuffled before annotation. We also monitored the process and discussed disagreements to minimize errors.

C.3 Extra Human Evaluation Results

In this section, we add additional human evaluation analysis to validate the performance of our proposed model, including *ET* domain to *VG* domain, *HK* domain to *SO* domain, and *HK* domain to *VG* domain. As shown in Table 6, Table 7 and Table 8, we have the following observations:

• All results across multiple low-resource scenarios demonstrated the effectiveness of our model for *SubjPQA*. Compared to the strong baselines of *Pegasus* and *InstructDS*, our improvements were consistent and significant, especially in the sparse 1% training data setting.

• In terms of the objective metrics of *Fact* and *Acc*, the proposed model achieved substantially higher scores. For example, with 1% target training data, the *Acc* score improved by 0.08 $\sim$ 0.11 over *InstructDS*. This indicated the model's ability to generate accurate objective answers by effectively transferring domain-invariant knowledge from rich source domains. That reflected the efficacy of the cross-domain knowledge adaptation.

• Similarly, for the subjective answers metrics of *Comp* and *Tcp*, the proposed model outperformed baselines by a large margin. Even with only 1% target data, the comprehension score surpassed *InstructDS* by 0.05 $\sim$ 0.12, showing the model's capacity to produce comprehensive subjective answers covering multi-perspective viewpoints. The significant boost in *Tcp* demonstrated the effectiveness of reinforcement learning.

• Moreover, consistent trends could be observed across different low-resource domain pairs like *Electronics-Sports*, *Electronics-Video Games*, etc. That indicated the robustness of our model.

Table 5: Comparisons of all the evaluated methods on different low-resource scenarios. Here, we use the HK domain as the source domain and VG as the low-resource domain. T-test, p-value<0.005

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$	$B_2 \uparrow$	$R_L \uparrow$
BM25	11.34	26.55	11.37	26.68	11.49	26.83	2.74	12.35	2.61	12.30	2.78	12.48
LexRank	11.51	26.81	11.62	26.94	11.71	27.05	2.95	12.64	2.87	12.57	2.98	12.71
PGNet	3.41	8.61	3.49	8.77	3.58	8.91	3.51	7.99	3.62	8.11	3.79	8.41
BART	21.76	27.97	21.93	29.21	23.67	30.64	24.18	28.19	25.79	29.51	25.03	29.11
Pegasus	25.02	30.25	25.93	31.47	26.81	32.97	28.51	32.91	28.97	33.69	29.41	34.24
InstructDS	27.87	33.24	28.41	34.62	29.17	35.21	45.92	49.36	47.12	50.63	48.23	51.72
PlanSum	-	-	-	-	-	-	4.72	12.49	4.81	12.68	4.93	12.79
FewSum	-	-	-	-	-	-	6.72	14.97	6.85	15.12	6.98	15.24
Ours	**31.92**	**38.69**	**32.61**	**39.41**	**33.81**	**40.21**	**52.11**	**56.27**	**53.49**	**57.41**	**54.72**	**58.63**

Table 6: Human evaluation results on different low-resource scenarios (source domain: ET; target domain: VG). T-test, p-value<0.005.

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑
Pegasus	0.55	0.57	0.59	0.60	0.62	0.61	0.48	0.69	0.51	0.72	0.53	0.73
InstructDS	0.60	0.61	0.63	0.63	0.65	0.66	0.63	0.83	0.64	0.85	0.66	0.86
Ours	**0.70**	**0.69**	**0.72**	**0.72**	**0.75**	**0.76**	**0.74**	**0.88**	**0.76**	**0.90**	**0.78**	**0.90**

Table 7: Human evaluation results on different low-resource scenarios (source domain: HK; target domain: SO). T-test, p-value<0.005.

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑
Pegasus	0.52	0.54	0.56	0.58	0.59	0.60	0.46	0.67	0.49	0.70	0.51	0.71
InstructDS	0.57	0.59	0.60	0.62	0.63	0.65	0.61	0.81	0.61	0.83	0.63	0.84
Ours	**0.68**	**0.67**	**0.70**	**0.70**	**0.73**	**0.74**	**0.72**	**0.86**	**0.74**	**0.88**	**0.76**	**0.88**

Table 8: Human evaluation results on different low-resource scenarios (source domain: HK; target domain: VG). T-test, p-value<0.005.

Scenarios	**Objective** Part of the Inductive Answer						**Subjective** Part of the Inductive Answer					
	Min (1%)		Med (5%)		Max (10%)		Min (1%)		Med (5%)		Max (10%)	
	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Fact ↑	Acc ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑	Comp ↑	Tcp ↑
Pegasus	0.54	0.56	0.58	0.60	0.61	0.62	0.47	0.68	0.50	0.71	0.52	0.72
InstructDS	0.59	0.60	0.61	0.64	0.63	0.65	0.61	0.82	0.62	0.84	0.64	0.85
Ours	**0.69**	**0.68**	**0.71**	**0.71**	**0.74**	**0.75**	**0.73**	**0.87**	**0.75**	**0.89**	**0.77**	**0.89**