Multi-Objective Forward Reasoning and Multi-Reward Backward Refinement for Product Review Summarization

Libo Sun[1], **Siyuan Wang**[1], **Meng Han**[2],
Ruofei Lai[2], **Xinyu Zhang**[2], **Xuanjing Huang**[3], **Zhongyu Wei**[1,4]
[1]School of Data Science, Fudan University, China,
[2]Huawei Poisson Lab, China
[3]School of Computer Science, Fudan University, China
[4]Research Institute of Intelligent Complex Systems, Fudan University, China,
lbsun23@m.fudan.edu.cn {wangsy18, xjhuang, zywei}@fudan.edu.cn
{hanmeng12, lairuofei, zhangxinyu35}@huawei.com

Abstract

Product review summarization aims to generate a concise summary based on product reviews to facilitate purchasing decisions. This intricate task gives rise to three challenges in existing work: factual accuracy, aspect comprehensiveness, and content relevance. In this paper, we first propose a **FB-Thinker** framework to improve the summarization ability of LLMs with multi-objective forward reasoning and multi-reward backward refinement. To enable LLM with these dual capabilities, we present two Chinese product review summarization datasets, **Product-CSum** and **Product-CSum-Cross**, for both instruction-tuning and cross-domain evaluation. Specifically, these datasets are collected via GPT-assisted manual annotations from an online forum and public datasets. We further design an evaluation mechanism **Product-Eval**, integrating both automatic and human evaluation across multiple dimensions for product summarization. Experimental results show the competitiveness and generalizability of our proposed framework in the product review summarization tasks.

Keywords: product review summarization, forward reasoning, backward refinement

1. Introduction

With the rapid growth of e-commerce platforms and consumer forums, customer-generated product reviews have become pivotal in commerce. These reviews offer invaluable perspectives on product quality and attributes, guiding consumer purchase decisions and also furnishing manufacturers with crucial product feedback. However, extracting meaningful insights from these massive reviews can be time-consuming and labor-intensive. This motivated a surge of research in the automatic product review summarization task (Boorugu and Ramesh, 2020; Brazinskas et al., 2021, 2020), aiming to condense extensive review content into coherent, concise, and self-consistent summaries.

While recent large language models (LLMs) with versatile generative capabilities have significantly improved general summarization (Tam et al., 2023; Hua et al., 2023), product review summarization remains intricate. As shown in Figure 1, it goes beyond mining key points but requires aggregation of diverse product aspects and collation of polarities, which poses three primary challenges. (1) Factual Accuracy: The summary must accurately depict the aspects and polarities of products as in the reviews; (2) Aspect Comprehensiveness: The summary needs to cover all aspects and features of products in the reviews, avoiding biases that solely focus on either strengths or weaknesses; (3) Content Relevance: The summary should remain concise, omitting any extraneous information unrelated to the products.

To improve these three dimensions for product review summarization, we propose a Forward-Backward Thinker (**FB-Thinker**) framework drawing inspiration from (Russell and Norvig, 2010). It empowers LLMs with both multi-objective forward reasoning and multi-reward backward refinement abilities. In the forward direction, the model can step-by-step deduce multiple objectives concerning aspects and polarities associated with each supporting sentence. This process facilitates the creation of accurate, comprehensive, and relevant summaries. In the backward direction, it can gather external feedback from the three reward models specified in the aforementioned dimensions, and accordingly refine the forward-generated summaries.

Specifically, we use LLaMA-7b (Touvron et al., 2023) as the backbone model and fine-tune it with our constructed instruction dataset to enhance its summarization ability for product reviews. To foster forward reasoning, we leverage Chain-of-Thought Prompting (Wei et al., 2022) and expand the fine-tuned datasets into a forward instruction set. It encompasses input reviews, output summaries, and multiple reasoning information including aspects, polarities, and corresponding supporting sentences. For backward refinement, we first build

Corpus:

The most dissatisfaction is the cost-effectiveness **(Low Costeffectiveness),** *the cheaper Qin models accelerate much faster than this car. There are also a lot of inconveniences caused by batteries. I have to charge it almost every two days, and it takes about 50 minutes to fully charge it, which makes my entire car use experience less than pleasant* **(Unpleasant using experience).** *Let's talk about BYD's after-sales service. I contacted it twice, once to install the battery guard, and once to clean the fender. Both times, I felt that the after-sales service was very good* **(Good Service),** *with return visits and good service attitude.*

Factual Inaccuracy:

The advantage of this car is that ***it is cost-effective*** *and has good after-sales service. The disadvantage is that the frequent charging makes using the car unpleasant.*

(Contrary to the low cost performance in the original corpus) ✗

Aspect Missing:

The disadvantages of this car are that, the power is not as good as cheaper models, and the car experience is unpleasant due to frequent charging.

(Good service Missing) ✗

Ground Truth:

The advantage of this car is that it has good after-sales service. The disadvantage is that the power is not as good as cheaper models, the cost-effective is not high, and because it requires frequent charging, the driving experience is not pleasant. ✓

Irrelevant Information:

The advantage of this car is good after-sales service, but the disadvantage is that ***it is noisy at high speeds****, and the frequent charging makes using the car unpleasant.*

(Noise issues were not mentioned in the original corpus) ✗

Figure 1: Examples of three mistakes in product review summarization: factual inaccuracies, irrelevant information and aspect missing.

three BERT-based reward models to provide feedback on forward-generated summaries in terms of accuracy, comprehensiveness, and relevance. Considering the unstable optimization challenges of multi-reward reinforcement learning (Christiano et al., 2017; Xue et al., 2023), we advocate fine-tuning the model to learn the refinement ability. We curate another backward instruction set for refining flawed summaries based on feedback. Overall, we fine-tune LLaMA-7b using both forward and backward instruction sets, bestowing it with the dual capability to reason forward and refine backward.

To explore product review summarization within Chinese contexts, we construct the first Chinese summarization dataset **Product-CSum** for automobile products. This dataset is sourced from the DCar website, a Chinese online forum dedicated to automobiles, and is split for both instruction tuning and evaluation purposes. Additionally, we present a benchmark **Product-CSum-Cross** covering three categories: restaurants, cosmetics, and automobiles, derived from publicly accessible datasets of aspect category sentiment analysis (Bu et al., 2021). This benchmark is designed to assess the general performance of our framework across diverse product domains.

Finally, we propose a novel evaluation mechanism **Product-Eval** for product review summarization. In addition to typical metrics in summarization tasks including BLEU, ROUGE, and METEOR, our evaluation incorporates three specially designed metrics for product summaries: factuality, comprehensiveness, and relevance. Both automatic and human evaluations are conducted in these three dimensions. For the human evaluation, we engage annotators to rate summaries across the three dimensions and also overall compare our summaries with other models. Results demonstrate the effectiveness of our FB-Thinker on both Product-CSum and Product-CSum-Cross datasets.

In general, our contributions are of three-folds:

- We present the pioneering Chinese product review summarization dataset, **Product-CSum**, and a cross-domain evaluation benchmark, **Product-CSum-Cross**[1].
- We propose a product review summarization framework, **FB-Thinker**. This framework is capable of generating high-quality summaries by forward reasoning and backward refinement.
- We design a novel evaluator, **Product-Eval**, which comprehensively assesses generated summaries and aligns with human evaluation.

2. Dataset

2.1. Task Formulation

The product review summarization task takes two primary inputs: a topic and an original review. (1) The topic, denoted as τ, specifies the product category or model to be summarized. (2) The review, denoted as x, comprises one or more consumer review sentences about the topic. The objective of this task is to generate a summary y about topic τ based on the review x, which should be accurate, comprehensive, and concise.

[1] https://github.com/sunlibo2390/Product-CSum-Cross-Dataset

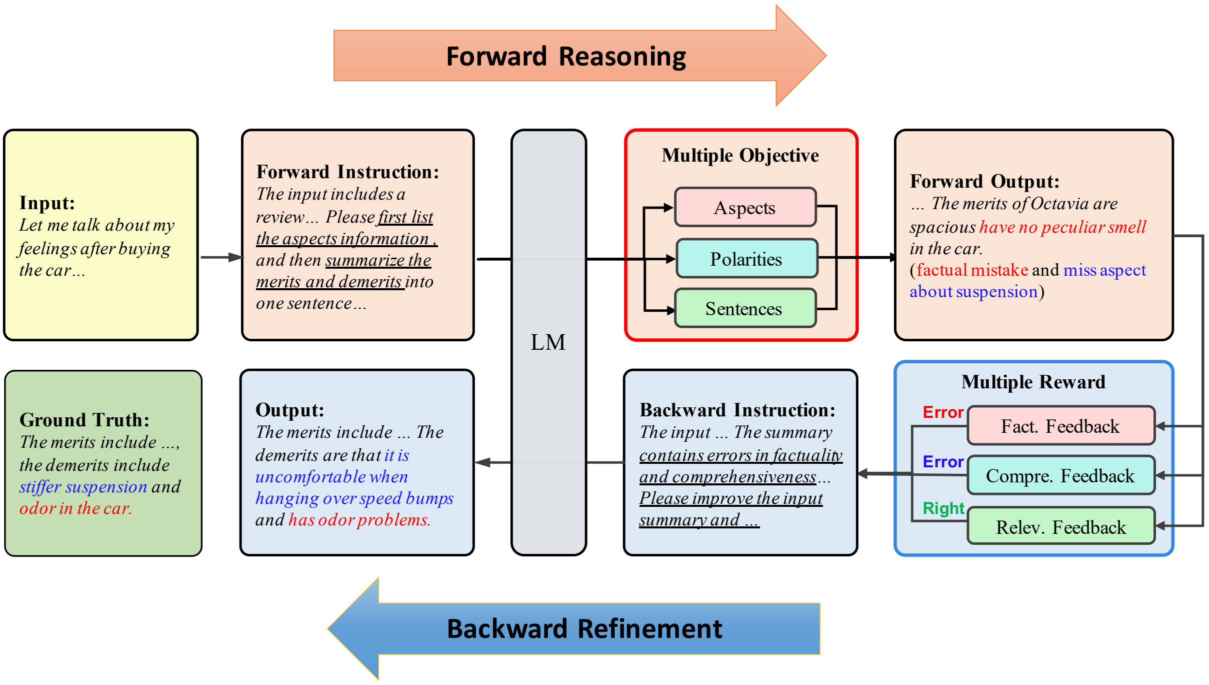
Forward Reasoning
Input:
Let me talk about my feelings after buying the car...
Forward Instruction:
The input includes a review... Please first list the aspects information, and then summarize the merits and demerits into one sentence...
LM
Multiple Objective
Aspects
Polarities
Sentences
Forward Output:
... The merits of Octavia are spacious have no peculiar smell in the car.
(factual mistake and miss aspect about suspension)
Ground Truth:
The merits include ..., the demerits include stiffer suspension and odor in the car.
Output:
The merits include ... The demerits are that it is uncomfortable when hanging over speed bumps and has odor problems.
Backward Instruction:
The input ... The summary contains errors in factuality and comprehensiveness... Please improve the input summary and ...
Multiple Reward
Error
Fact. Feedback
Error
Compre. Feedback
Right
Relev. Feedback
Backward Refinement

Figure 2: The overview of our proposed framework FB-Thinker. In the forward phase, forward instruction guides LLMs to generate a candidate summary with multiple intermediate reasoning objectives. In the backward phase, three pre-trained reward models provide feedback on potential errors within this candidate summary, and backward instruction directs LLMs to rectify these errors and refine the summary.

2.2. Dataset Creation

Our Chinese product review summarization dataset, **Product-CSum**, originates from the DCar [2] website, a Chinese online automobile forum. It encourages users to post and discuss their reviews of vehicle experiences, stimulating potential buyers. Additionally, our cross-domain evaluation benchmark, **Product-CSum-Cross**, is sourced from public datasets of aspect category sentiment analysis (ACSA). These datasets feature product reviews coupled with corresponding aspects and polarities, spanning three domains: restaurant (Bu et al., 2021), cosmetics [3] and automobiles [4].

For both datasets, we follow three annotation steps. First, we extracted the topic and the review from each sample to serve as the input. Second, we employed ChatGPT [5] with in-context learning to generate candidate summaries. Finally, human annotators were asked to verify whether the summaries generated by ChatGPT were correct. The remaining summaries after filtering are deemed as the ground truth output for each instance.

We recruited 56 annotators, each possessing at least a bachelor's degree. From this endeavor, we accumulated 6, 939 review texts paired with their corresponding topics, and a total of 27, 241 summaries. The detailed statistics of Product-CSum and Product-CSum-Cross are listed in Table 1.

	Product-CSum-Cross			Product-CSum
	Cos.	Auto.	Res.	Auto.
# of reviews	1,287	1,142	840	3,659
# of summary	4,786	4,969	3,572	1,2113
Avg. # words				
per review	152.9	272.2	276.4	643.7
per summary	60.8	69.2	60.7	62.8

Table 1: The statistics of datasets. "Cos.", "Auto." and "Res." are respectively abbreviations of cosmetics, automobiles and restaurants.

3. FB-Thinker

Figure 2 illustrates the overall framework of our FB-Thinker. It is mainly composed of two parts: 1) Multi-Objective Forward Reasoning: empowering LLMs with the capability to craft product summaries through intermediate reasoning on detailed aspects, polarities, and supporting sentences, and 2) Multi-Reward Backward Refinement: learning

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

[3] https://github.com/xmxoxo/Text-Opinion-Mining

[4] https://www.datafountain.cn/competitions/310/datasets

[5] Specifically, we use gpt-3.5-turbo as in https://platform.openai.com/docs/models/gpt-3-5, and ditto.

to provide feedback using three reward models for identifying summary errors, and refine the imperfect summaries. We first introduce the base instruction-tuning process to endow the model with a tailored product review summarization ability (Sec. 3.1), and then introduce these two modules (Sec. 3.2 & Sec. 3.3). We finally introduce the training and inference procedures(Sec. 3.4).

3.1. Product Review Summarization Learning

To develop the LLMs' versatile generative ability towards specific product review summaries, we first construct a one-shot instruction dataset based on Product-CSum. We then fine-tune the model using this curated dataset.

Instruction Set Creation We construct the instruction-tuning dataset as "instruction-response" pairs. The response is the ground truth summary. The instruction encompasses both a task-defining prompt and a product review to be summarized. Considering that in-context learning can enhance the model's instruction-following performance (Dong et al., 2022), we adopt a one-shot prompting approach and incorporate a reference demonstration within each instruction. An illustration of our one-shot instruction is as follows:

The input includes a review text about [topic]. Please refer to the following example, and compile a summary that outlines both the advantages and disadvantages associated with [topic] based on the input text.

## Example
Input: [Reference Review]
Output: [Reference Summary]

We finally obtained 20, 628 samples in our one-shot instruction set, using carefully prepared examples.

Low-Rank Tuning Utilizing the above one-shot instruction set, we fine-tuned the Chinese LLaMA-7b model with LoRA (Lower-dimension Optimization for Robust Attention) technique. LoRA maps the weight update within the self-attention module's projection matrices in the Transformer architecture to a lower-dimensional space, followed by reverting to the original output dimension. In our work, we applied LoRA to all Query/Key/Value/Output projection matrices within the self-attention module.

3.2. Multi-Objective Forward Reasoning

To encourage the model to generate more accurate, comprehensive, and relevant product summaries, we adopt the CoT prompting strategy and enrich the one-shot instruction with intermediate reasoning details of product aspects, polarities, and supporting sentences. We further fine-tune the model using this forward instruction set to gain the forward reasoning ability.

Forward Instruction Set We utilize ChatGPT to extract aspects and corresponding polarities within each context, in the format of *aspect-polarity-supporting sentence*. We modify both instructions and responses in the one-shot instruction set. The forward instructions aim to guide the model to first enumerate the multiple aspect information (including aspects, polarities, and supporting sentences) in the review, and subsequently generate a summary. Regarding the responses, we add aspect information as the intermediate reasoning process prior to the output summary. Similarly, the reference examples in the instructions are also expanded with aspect information. A sample is shown below:

The input includes a review text about [topic]. Please refer to the following example, first list all aspect information about [topic] based on the input text, and then compile a summary that outlines both the advantages and disadvantages associated with [topic].

## Example
Input: [Reference Review]
Output:
[Aspect1-Polarity1-Sentence1]
[Aspect2-Polarity2-Sentence2]
. . .
[Reference Summary]

3.3. Multi-Reward Backward Refinement

Our goal is to use multiple reward models to provide feedback on the generated summaries emphasizing accuracy, comprehensiveness, and relevance. Following identifying errors in these dimensions, we then train the model to refine the summaries based on the feedback.

Reward Models To achieve this, we first train three BERT-based reward models, each tailored for one of the aforementioned dimensions. These models aim to detect potential errors within the summaries concerning their respective dimensions. Thus, we need to construct datasets for training reward models, which contain both optimal summaries and imperfect ones with specific errors. Take the ground truth as optimal summaries, the imperfect samples with three types of errors are constructed as follows:

- **Incorrect** We use ChatGPT to alter the polarity of one or more aspects within the optimal summary, thereby synthesizing an incorrect summary.
- **Incomplete** We mask certain aspects and summarize the left information, or focus solely on

either advantages or disadvantages, thus generating an incomplete summary.

- **Irrelevant** We sample sentences from a sentence pool that are irrelevant to the original review. Inserting them into the optimal summary or merging them with the original review and then summarizing, yields a summary with irrelevant content.

Considering that multiple types of errors may occur at the same time, we also create mixed-error samples by arranging the three mentioned error types in a combinatorial manner. Utilizing these synthesized error samples along with the ground truth, we constructed three sets of classification data. Positive samples denote the occurrence of errors from the respective category (though other errors might also be present), while negative samples suggest the absence of such errors.

Backward Instruction Set We construct a backward instruction set to enable the model to have the backward refinement capability based on feedback through fine-tuning. Each instruction in this set comprises three key elements: a review text, a flawed summary containing errors, and a task-defining prompt that specifies the error types present in the imperfect summary. This instruction guides the model to fix the mistakes accordingly and generate a refined summary. The response is also the ground truth summary. Below is a backward instruction for only factual errors:

The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.

The summary contains errors in factual accuracy. Factual accuracy means that the polarity of the aspects contained in the original text is correctly judged and that the advantages and disadvantages are correctly classified. Please improve the input summary and output the improved summary.

Input: [Review] [Summary]

Given that a single summary may contain different combinations of errors, we craft different refinement instructions for all 7 distinct error type combinations, which are available at Appendix A.1.

3.4. Training & Inference

Overall, we simultaneously utilize forward and backward instruction sets to fine-tune the model to gain the dual capability of forward reasoning and backward refinement.

In the inference phase, our summarization pipeline consists of two stages: forward-reasoning and backward-refinement. During the forward-thinking, forward instruction directs the model to first list the detailed aspects, polarities, and supporting sentences within the input review before generating the summary. In the subsequent backward-refinement stage, three pre-trained reward models are employed to identify errors within the forward-generated summary. Guided by the feedback from these reward models, the model rectifies the errors, thereby generating an improved summary.

4. Experiments

4.1. Experimental Setup

Dataset Our experiments are conducted on the test set of Product-CSum. It consists of 1, 800 triplets in the format of <topic, review, summary>. Experiments are also performed on Product-CSum-Cross for cross-domain evaluation, which respectively contains 4, 786, 4, 969, and 3, 572 triplets in restaurant, cosmetic, and automobile scenarios.

Implementation Details When training the base LM with instruct-tuning, we combine the forward reasoning instruction set and backward refinement instruction set as the fine-tuned dataset. We use Chinese LLaMA-7b as the base model, set the learning rate to 3×10^{-4}, batch size to 16, gradient accumulation step to 4, and train the model 50 epochs on 8 NVIDIA RTX3090 GPUs. The α and r of the LoRA method are both set to 16. When training the reward models, We use BERT-base-Chinese (Devlin et al., 2018) as the base model. We set the learning rate to 1×10^{-5}, batch size to 16, and train on an NVIDIA RTX3090 GPU for 5 epochs. For training both models, we employed AdamW as the optimizer.

Model for Comparison We compare our FB-Thinker framework with several baselines, including both fine-tuned models and off-the-shelf ones:

- BART (Lewis et al., 2019): we fine-tuned BART using the training set of Product-CSum to adapt to this task.
- LLaMA (Touvron et al., 2023): We chose the Chinese LLaMA-7b and compared two variants with and without fine-tuning using our one-shot instruction dataset.
- Alpaca-LoRA (Taori et al., 2023): a model obtained by LoRA-tuning LLaMA-7b based on the Alpaca instruction set.
- Baichuan (Yang et al., 2023): a language model trained in English and Chinese, and we choose Baichuan-7b as our compared model.

4.2. Evaluation Metrics

Our evaluation metrics include automatic evaluation metrics and human evaluation.

	Model	BLEU	ROUGE	METEOR	Fact.	Compre.	Relev.	# Param (b)
w/ finetuning	BART	17.78	20.60	15.50	**82.91**	77.33	17.62	0.1
	LLaMA	36.52	36.42	32.32	54.49	88.17	29.42	7
w/o finetuning	Baichuan	17.58	18.84	12.95	62.87	80.78	11.73	7
	LLaMA	24.19	26.58	20.24	72.64	84.83	19.01	7
	Alpaca-LoRA	29.77	30.91	25.71	63.16	86.11	23.49	7
Ours	**FB-Thinker**	**41.38**	**40.81**	**38.98**	67.69	**89.00**	**37.11**	7

Table 2: Main result on product reviews summarization on Product- CSum test set. We report BLEU-1 (BLEU), ROUGE-L (ROUGE), METEOR, Factuality (Fact.), Comprehensiveness (Compre.), Relevance (Relev.) and the number of parameters in each model (# Param).

Automatic evaluation First, we adopt typical n-gram evaluation metrics for summarization, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005).

In addition, to assess performance in factual accuracy, aspect comprehensiveness, and content relevance, we introduce three novel metrics: Factuality, Comprehensiveness, and Relevance. Rather than directly using three reward models for evaluation which hinges on the BERT performance, we rely on ChatGPT and design heuristic rules to yield more objective results.

- **Factuality**: We use ChatGPT to extract the aspects and polarities from both the review and the generated summary. The factuality scores only if all aspect polarities exactly align with the review.
- **Comprehensiveness**: Considering an aspect might be expressed in various ways in the summary, we use ChatGPT to extract the supporting sentences of each aspect and compare our summary with them. If any word overlaps after removing stop words, we label the aspect as "covered". Otherwise, it's labeled "missing".
- **Relevance**: We use Intersection over Union (IoU) (Rezatofighi et al., 2019), to compare the generated summary with the ground truth, after text segmenting and removing stop words.

Human Evaluation To assess the summary quality aligning with human value, we randomly selected 50 samples for human evaluation using both rating and comparison methods.

In the rating mechanism, for each instance, human annotators are asked to read the review with its topic and rate the summary on a scale of 1 (worst) to 3 (best) across three dimensions: (1) Factuality: Assessing the accuracy of all aspect polarities presented in the summary. (2) Comprehensiveness: Evaluating whether the summary adequately covers all aspects discussed in the review. (3) Relevance: Determining if the summary includes irrelevant information or is overly verbose.

In the comparison method, we ask annotators to compare the summaries generated by ChatGPT (ground truth), FB-Thinker, and fine-tuned LLaMa, and choose the best one in overall quality. If overall scores are tied, we successively compare based on factuality, comprehensiveness, and relevance.

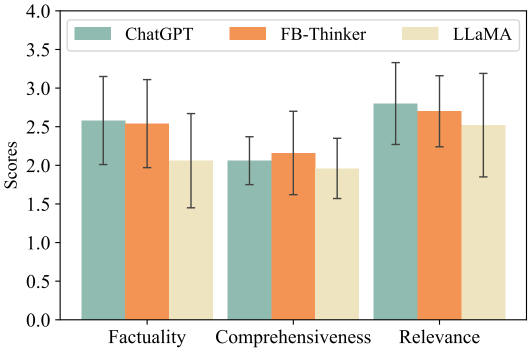

Figure 3: Result of human evaluation on facticity, relevance, and comprehensiveness. LLaMA means its fine-tuned variant, the same as below.

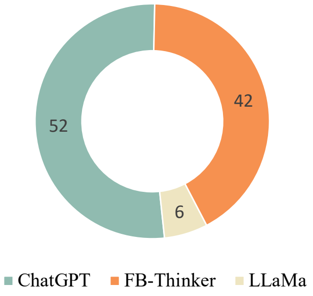

Figure 4: Overall Comparison Result. Numbers are the ratio of system output chosen as the best.

4.3. Main Results

We report the evaluation results on the Product-CSum test set for assessing in-domain product review summarization.

Automatic Evaluation As the results listed in Table 2, we have the following findings:

Model	BLEU	ROUGE	METEOR	Fact.	Compre.	Relev.
Component Ablation						
FB-Thinker	41.38	**40.81**	**38.98**	**67.69**	89.00	**37.11**
- Forward Reasoning	**41.61**	40.69	38.27	67.10	**89.39**	35.93
- Backward Refinement	40.42	39.91	38.09	62.66	89.06	36.32
- Forward & Backward Thinking	36.52	36.42	32.32	54.49	88.17	29.42
Reward Ablation						
FB-Thinker	41.38	**40.81**	**38.98**	**67.69**	89.00	**37.11**
only accuracy feedback	41.17	40.66	38.80	67.26	**89.17**	36.95
only comprehensiveness feedback	41.29	40.72	38.89	67.60	89.11	37.02
only relevance feedback	41.31	40.76	38.91	67.53	89.17	37.06

Table 3: Ablation analysis of different components and feedback rewards in our FB-Thinker framework.

Domain	Model	BLEU	ROUGE	METEOR	Fact.	Compre.	Relev.
Cosmetic	Alpaca-LoRA	32.16	32.05	26.33	**36.67**	52.29	24.80
	Ours	**33.73**	**32.47**	**27.87**	36.21	**54.93**	**26.17**
Automobile	Alpaca-LoRA	32.16	32.85	27.83	**48.34**	58.93	25.10
	Ours	**34.00**	**34.19**	**29.77**	46.23	**82.92**	**26.76**
Restaurant	Alpaca-LoRA	26.89	27.96	23.38	35.26	60.17	21.82
	Ours	**28.88**	**29.16**	**24.93**	**35.98**	**62.21**	**23.40**

Table 4: Experimental results on Product-CSum-Cross dataset.

- Our FB-Thinker significantly outperforms all baselines across most metrics. This indicates the effectiveness of our method with forward reasoning and backward refinement.
- The fine-tuned LLaMA generates substantial improvements over other baselines, especially LLMs without fine-tuning. This observation highlights the value of instruction tuning within specific domains, directing general LLMs toward product review summarization ability.
- The high factuality scores of BART and LLaMa (w/o fine-tuning) result from their limited aspect comprehensiveness. Missing aspects would not be used to evaluate factuality and lead to misleadingly high factuality scores.

Human Evaluation We report the rating and comparison result of human evaluation in Figure 3 and Figure 4. We have the following observations:

- Fine-grained results show that our model generates more accurate, comprehensive, and relevant summaries than fine-tuned LLaMA. While it matches ChatGPT in factuality and relevance, it even surpasses in comprehensiveness.
- In an overall comparison, our approach significantly outperforms the fine-tuned LLaMA and achieves comparable preference to powerful ChatGPT which is our ground truth.

These results reveal high human preferences for our generated summaries via forward reasoning and backward refinement.

4.4. Ablation Study

We conduct ablation studies to explore the effectiveness of different components and different feedback rewards in FB-Thinker framework on Product-CSum test set.

Component Ablation We compare our FB-Thinker against three variants for component ablation. 1) Without forward reasoning: we use the backward instruction set and replace the forward set with the original one-shot set. 2) Without backward refinement: we remove the backward instruction set of the original framework to fine-tune the model for solely forward reasoning ability. 3) Without either forward or backward thinking: we remove both the forward and backward instruction sets and only use the original one-shot instruction set for fine-tuning. Results are presented in the upper part of Table 3. Both forward reasoning and backward refinement significantly improve summary quality, whether used individually or in parallel.

Reward Ablation We further investigate the impact of feedback from three reward models. During the backward refinement phase, we respectively utilize one of three reward models (accuracy, comprehensiveness, and relevance) to provide a single feedback for backward improvement. The results are listed in the lower part of Table 3. Compared to the overall framework, gathering feedback from just one reward model results in a notable reduction in model performance. This reveals the effectiveness

Review: Although this A6L is low-power, 95% of it is used in urban areas. Compared with the 1.6T 408, the back feeling is not as obvious as the 408... The power is slightly worse than the Q5 7.8 years ago... In terms of space, it's really uncomfortable to sit in the back row. To be honest, the trunk is not too big... In terms of the interior, there is no place to put mobile phones... The armrest box is as shallow as anything, and can only hold an ID card and so on... Let's summarize the advantages. The appearance is not stunning, but you will find that it is more and more attractive, the price is right, the power is sufficient, and the interior is not cheap, suitable for business and IKEA... The disadvantage is that you must pay attention to the gap between the front windshield and the instrument panel, a lighter may fall in ...
Ground Truth: The advantages of Audi A6L are its attractive appearance, affordable price, sufficient power, not cheap interior, and suitable for shopping in IKEA. However, it has disadvantages such as not being as powerful as the 408 and the 7.8-year-old Q5 of the same level, higher fuel consumption, uncomfortable rear seats, and insufficient storage space.
FB-Thinker: The advantages of the Audi A6L are its good power performance when used in urban areas, its affordable price, good interior quality, and attractive appearance. However, there are disadvantages such as the gap between the front windshield and the instrument panel, which is easy to fall into, the posture of the rear seats is uncomfortable, and the fuel consumption is high.
w/o Forward-Reasoning: The advantage of the Audi A6L is that it has sufficient power, which is 95% when used in urban areas. Its disadvantages include higher fuel consumption, poor rear seat comfort, smaller space, unreasonable interior design, and safety issues.
w/o Backward-Refinement: The advantages of the Audi A6L are its strong power, obvious push-back feel, and attractive appearance.
Alpaca-LoRA: The advantage of BYD is that the rear space is spacious and comfortable, and the after-sales service is good. The disadvantage is that the main driver and trunk space are too small for the author, and due to the need for frequent charging, the driving experience is not pleasant.

Table 5: Case study of an instance in the test set.

of feedback derived from our built reward models.

4.5. Cross-domain Generalizability

To further investigate the cross-domain generalizability, we test on the Product-CSum-Cross benchmark, covering restaurant, cosmetic, and automobile domains. We compare our FB-Thinker with Alpaca-LoRA, the top-performing off-the-shelf LLMs among our baselines. The results are shown in the Table 4. Our framework consistently demonstrates superior performance across most metrics within all domains. This indicates its cross-domain generalization ability to generate high-quality summaries for specific domains, even if they are not part of the training data.

4.6. Case study

We illustrate the superiority of our model through a case study in Table 5. Alpaca-LoRA struggles to summarize the pros and cons of A6L, which mostly replicates reference instructions. In contrast, our FB-Thinker generates a summary that closely aligns with the ground truth, offering an accurate, comprehensive and relevant outline. Without incorporating forward reasoning or backward refinement, our framework risks missing crucial aspects such as “appearance” and “price”. It might even generate only the advantages or disadvantages, rather than a balanced view. Moreover, our FB-Thinker can generate more accurate expressions on “power performance” with specific constraints than the ground truth and other models.

5. Related Work

Product Review Summarization Product review summarization aims to condense product reviews into brief summaries to facilitate consumers to swiftly access product information. Despite several available datasets (Brazinskas et al., 2021, 2020; Angelidis et al., 2021) for this task, the lack of Chinese datasets restricts the review summarization capabilities in Chinese contexts. Previous research on product review summarization mainly falls into two categories: supervised and unsupervised approaches. Supervised methods typically involve fine-tuning pre-trained language models (Isonuma et al., 2017; Oved and Levy, 2021b; Xu et al., 2023), using manually curated product summaries within specific domains. However, these methods are highly domain-sensitive, limiting their capabilities of generalization to diverse scenarios. On the other hand, unsupervised techniques (Zhao and Chaturvedi, 2020; Li et al., 2023) aims to identify significant aspects within reviews by employing predefined aspect seed words or linguistic features, which can be susceptible to data noise.

Summarization with LLMs Large language models (LLMs) with advanced generative capabilities have demonstrated remarkable performance in various summarization tasks. Zhang et al. (2023) assess the performance of various LLMs using a meticulously curated, human-annotated news summarization dataset. Ding and Ito (2023) propose a novel framework for fine-tuning LLMs to autonomously identify points of agreement among diverse opinions. As demonstrated by Ahmed and Devanbu (2022), LLMs are equipped with few-shot training to enhance code summarization capabilities. To further optimize the summarization performance, Madaan et al. (2024) prompt LLMs to refine their generation results. Lastly, Gou et al. (2023) allows static LLMs to validate and self-correct their output by interacting with external tools, enhancing their reliability and accuracy.

6. Conclusions

In this paper, we first introduce two product review summarization datasets in Chinese contexts, Product-CSum and Product-CSum-Cross for instruction-tuning and evaluation. Specifically, we collect original corpus from DCar automobile forum and public ACSA datasets and perform manual annotation after the preprocessing by ChatGPT. To address the three challenges: factual accuracy, aspect comprehensiveness, and content relevance posed by product review summarization, we further construct a FB-Thinker framework with multi-objective forward reasoning and multi-reward backward refinement. We additionally propose three metrics to measure the performance of summarization results in the dimensions of factuality, comprehensiveness, and relevance. Experiments demonstrate that our framework competes favorably with the comparison models.

7. Ethics Statement

Our data collection is both ethically sound and legally compliant. The information gathered from DCar is publicly accessible and devoid of any sensitive or private data. We cited the three publicly available ACSA datasets used in our research paper. Furthermore, we engaged 56 part-time annotators, remunerating them at a minimum hourly compensation of at least $3.9 which surpasses the local minimum wage.

8. Acknowledgments

This work is supported by National Natural Science Foundation of China (No.71991471, No. 6217020551) and Science and Technology Commission of Shanghai Municipality Grant (No.21DZ1201402).

9. Bibliographical References

T. Ahmed and P. Devanbu. 2022. Few-shot training llms for project-specific code-summarization. *In Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, Online:1–5.

S. Angelidis, R.K. Amplayo, Y. Suhara, X. Wang, and M. Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

R. Boorugu and G. Ramesh. 2020. A survey on nlp based text summarization for summarizing product reviews. *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Online:352–356.

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

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, Online:9424–9442.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. Asap: A chinese review dataset towards aspect category sentiment analysis and rating prediction. *arXiv preprint arXiv:2103.06605*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shiyao Ding and Takayuki Ito. 2023. Self-agreement: A framework for fine-tuning language models to find agreement among diverse opinions. *arXiv preprint arXiv:2305.11460*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint*, arXiv:2305.11738.

Y. Hua, Z. Deng, and K. Mckeown. 2023. Improving long dialogue summarization with semantic graph representation. *Findings of the Association for Computational Linguistics: ACL 2023*, Online:13851–13883.

Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. 2017. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on empirical methods in natural language processing*, pages 2101–2110.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

H. Li, S.B.R. Chowdhury, and S. Chaturvedi. 2023. Aspect-aware unsupervised extractive opinion summarization. *Findings of the Association for Computational Linguistics: ACL 2023*, Online:12662–12678.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

N. Oved and R. Levy. 2021a. Pass: Perturb-and-select summarizer for product reviews. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Volume 1: Long Papers:351–365.

Nadav Oved and Ran Levy. 2021b. Pass: Perturb-and-select summarizer for product reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 351–365.

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.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Stuart J Russell and Peter Norvig. 2010. *Artificial intelligence a modern approach*. London.

Tejpalsingh Siledar, Jigar Makwana, and Pushpak Bhattacharyya. 2023. Aspect-sentiment-based opinion summarization using multiple information sources. pages 55–61.

D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. *Findings of the Association for Computational Linguistics: ACL 2023*, Online:5220–5255.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

H. Xu, H. Liu, Z. Lv, Q. Yang, and W. Wang. 2023. Pre-trained personalized review summarization with effective salience estimation. *Findings of the Association for Computational Linguistics: ACL 2023*, Online:10743–10754.

Wanqi Xue, Bo An, Shuicheng Yan, and Zhongwen Xu. 2023. Reinforcement learning from diverse human preferences. *arXiv preprint arXiv:2301.11774*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T.B. Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint*, arXiv:2301.13848.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9644–9651.

A. Appendices

A.1. Refinement instructions

Error Type	Prompt Template
Factual Accuracy	*The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.* *The summary contains errors in factual accuracy. Factual accuracy refers to correctly judging the polarity of aspects mentioned in the original text, as well as accurately classifying the advantages and disadvantages. Please enhance the input summary and provide the improved version.* *Input: [Review] [Summary]*
Aspect Comprehensiveness	*The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.* *The summary contains errors in aspect comprehensiveness. Aspect comprehensiveness requires that the summary includes all aspect information mentioned in the original text, ensuring that both advantages and disadvantages are considered. Please enhance the input summary and provide the improved version.* *Input: [Review] [Summary]*
Content Relevance	*The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.* *The summary contains errors in content relevance. Content relevance means that the summary only includes relevant information about {topic} and does not include other irrelevant information. Please enhance the input summary and provide the improved version.* *Input: [Review] [Summary]*
Factual Accuracy & Aspect Comprehensiveness	*The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.* *The summary contains errors in factual accuracy and aspect comprehensiveness. Factual accuracy refers to correctly judging the polarity of aspects mentioned in the original text, as well as accurately classifying the advantages and disadvantages. Aspect comprehensiveness requires that the summary includes all aspect information mentioned in the original text, ensuring that both advantages and disadvantages are considered. Please enhance the input summary and provide the improved version.* *Input: [Review] [Summary]*
Aspect Comprehensiveness & Content Relevance	*The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.* *The summary contains errors in content relevance and content relevance. Aspect comprehensiveness requires that the summary includes all aspect information mentioned in the original text, ensuring that both advantages and disadvantages are considered. Content relevance means that the summary only includes relevant information about {topic} and does not include other irrelevant information. Please enhance the input summary and provide the improved version.* *Input: [Review] [Summary]*
Factual Accuracy & Content Relevance	*The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.* *The summary contains errors in factual accuracy and content relevance. Factual accuracy refers to correctly judging the polarity of aspects mentioned in the original text, as well as accurately classifying the advantages and disadvantages. Content relevance means that the summary only includes relevant information about {topic} and does not include other irrelevant information. Please enhance the input summary and provide the improved version.* *Input: [Review] [Summary]*
Factual Accuracy & Aspect Comprehensiveness & Content Relevance	*The input includes a review text about [topic] and a summary of the advantages and disadvantages of [topic] based on this text.* *The summary contains errors in factual accuracy and content relevance. Factual accuracy refers to correctly judging the polarity of aspects mentioned in the original text, as well as accurately classifying the advantages and disadvantages. Aspect comprehensiveness requires that the summary includes all aspect information mentioned in the original text, ensuring that both advantages and disadvantages are considered. Content relevance means that the summary only includes relevant information about {topic} and does not include other irrelevant information. Please enhance the input summary and provide the improved version.* *Input: [Review] [Summary]*