

InstructPTS: Instruction-Tuning LLMs for Product Title Summarization

Besnik Fetahu Zhiyu Chen Oleg Rokhlenko Shervin Malmasi
Amazon.com, Inc. Seattle, WA, USA
{besnikf,zhiyu,olegro,malmasi}@amazon.com

Abstract

E-commerce product catalogs contain billions of items. Most products have lengthy titles, as sellers pack them with product attributes to improve retrieval, and highlight key product aspects. This results in a gap between such unnatural products titles, and how customers refer to them. It also limits how e-commerce stores can use these seller-provided titles for recommendation, QA, or review summarization.

Inspired by recent work on instruction-tuned LLMs, we present InstructPTS, a controllable approach for the task of Product Title Summarization (PTS). Trained using a novel instruction fine-tuning strategy, our approach is able to summarize product titles according to various criteria (e.g. number of words in a summary, inclusion of specific phrases, etc.). Extensive evaluation on a real-world e-commerce catalog shows that compared to simple fine-tuning of LLMs, our proposed approach can generate more accurate product name summaries, with an improvement of over 14 and 8 BLEU and ROUGE points, respectively.

1 Introduction

E-commerce product catalogs (e.g. Amazon, Walmart) contain billions of products with lengthy names: 65% of product titles have more than 15 words (Rozen et al., 2021). This is due to sellers overloading titles with extra information about product functionality, colors, sizes and more in order to maximize their search rankings for as many queries as possible, and to captivate customers.

However, this can lead to poor experiences when these titles need to be used in other contexts such as being read aloud by voice assistants, referenced in narrative text such as product summaries, or rendered in text interfaces with limited display sizes.

This has resulted in the practical task of Product Title Summarization (PTS), which aims to extract a natural representation corresponding to how humans would refer to the product (Sun et al., 2018).

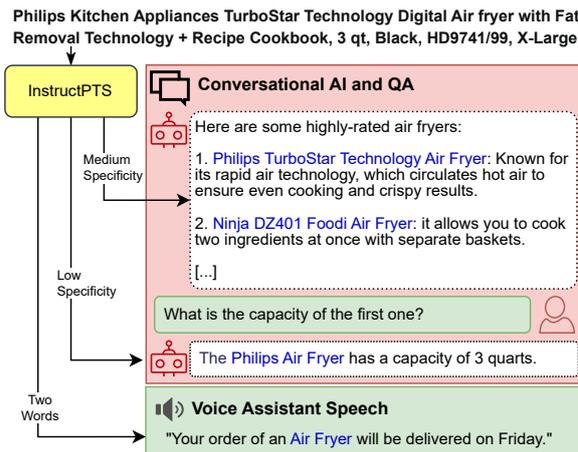


Figure 1: Example of how an original product title is reformulated by InstructPTS for different applications.

As shown by the example in Figure 1, these summarized titles can then be used in other tasks like voice assistant speech, product QA, summarization, recommendation, and query understanding.

Most work thus far has used traditional abstractive and extractive summarization methods to create a single summary. Inspired by recent advances in Large Language Models (LLMs) and instruction-tuning, we present InstructPTS, the first PTS approach to use instruction fine-tuning (IFT) of LLMs to achieve controllable title summarization across different dimensions such as: (i) desired length, (ii) presence of specific words (e.g. brands, size, etc.), and (iii) summary specificity. Figure 2 shows supported instructions, which capture various requirements, and are automatically generated from a parallel dataset of original product titles and summaries. A key advantage of InstructPTS is that it allows us to utilize a single model for generating multiple titles for different downstream tasks.

Evaluation on a leading real-world e-commerce catalog shows that our InstructPTS approach generates *accurate* summaries, and has high instruction-following capability. Furthermore, the generated

Item Name: “Blade Tail Rotor Hub Set B450 330X Fusion 270 BLH1669 Replacement Helicopter Parts”

- Summarize {Item_Name} to contain at most 3 words → “Blade Rotor Hub”
- Summarize {Item_Name} with Low specificity and to contain the words “B450 330X” → “Rotor Hub Set B450 330X”
- Summarize {Item_Name} with Low specificity → “Rotor Hub Set”

Figure 2: A sample of product title summaries generated by InstructPTS for different instructions.

summaries are judged by humans as being highly relevant and capturing the most salient words from the original title. Finally, extrinsic evaluation using a retrieval system shows that the summarized titles retain sufficient unique characteristics of the product to retrieve it with high accuracy.

2 Related Work

PTS falls within the broader domain of text summarization techniques (El-Kassas et al., 2021).

Both extractive and abstractive summarization approaches have been applied for PTS. For example, Wang et al. (2018) propose a multi-task learning framework, where one network summarizes the product name, while another learns to generate search queries. Sun et al. (2018) propose a multi-source pointer network to generate short product names from longer input names and background knowledge. Gong et al. (2019) developed an enhanced feature extraction approach to generate short product names by incorporating external word frequency information and named entities as additional features. An different approach based on Generative Adversarial Networks that encode multi-modality features (such as product images and attribute tags) is presented by Zhang et al. (2019). Xiao and Munro (2019) adopt Bi-LSTMs to extract key words for product name summaries. Subsequently, Mukherjee et al. (2020) tackled the vocabulary mismatch problem by integrating pre-trained embeddings with trainable character-level embeddings as inputs to Bi-LSTMs. An adversarial generation model that can generate personalized short names is proposed by Wang et al. (2020).

Our approach differs from prior work in two aspects. Firstly, previous studies primarily focused on generating a single product name summary, which may not cater to the diverse use cases in e-commerce applications. In contrast, our approach

offers the flexibility to generate diverse summary types (e.g. specific number of words, specific summary specificity etc.). Secondly, drawing inspiration from the recent success of LLMs (Ouyang et al., 2022; Longpre et al., 2023), we are the first to propose an instruction-based approach for PTS.

3 InstructPTS Approach

We now outline our proposed InstructPTS approach: we describe the base model, and provide details about the instruction fine-tuning.

3.1 Base Model

The base model for InstructPTS is FLAN-T5 (Chung et al., 2022), an LLM pre-trained on a large set of instruction fine-tuning tasks. We opt for this LLM family given that they are suitable for instruction fine-tuning (IFT) for our task. We experiment with different model sizes (cf. §4.2), and compare the advantage of IFT over other training strategies.

3.2 Ground Truth Dataset

We use a parallel dataset of original product title and summary pairs. The summaries are of two *specificity* levels: Low or Medium, which control how descriptive it is w.r.t. the original title. Low summaries are short (approx. 2 ($SD=\pm 1$) words) and typically do not include brand or other product details, but instead focus on a highly abstract description of the product family. Medium summaries are longer (approx. 4 ($SD=\pm 1.4$) words) and contain brand/model names, and aspects that identify the specific product. This gold data is generated using a hybrid approach: a sequence tagger chunks words that need to be included in the summary, and human annotators accept/reject the taggers decision, or rewrite the summary entirely. This is an extractive process; the summaries only contain words that appear in the original product title.

The data is split into train/dev/test sets with 100k/10k/1M product titles, respectively. Summaries of Medium specificity make up 58% of the data; the remaining 42% are of Low specificity. The same products can have both levels, but not always.

3.3 Instruction Fine-Tuning

LLM instruction fine-tuning (Ouyang et al., 2022) has proven to improve generalizability, allowing LLMs to perform better on tasks defined using natural. IFT allows LLMs to flexibly encode various constraints defined in natural language, enabling robust and controllable performance.

#	Instruction	Instruction Goal	Product Title (input)	Product Title Summary
1	Summarize {Item Name} with Low specificity	Specificity Constraints.	"EcoSafe 6400 Certified Compostable Bags 2.5 Gallon (16" x 17"), (Case of 360 Bags : 12 Rolls)"	Compostable Bags
2	Summarize {Item Name} with Medium specificity			EcoSafe Compostable Bags
3	Summarize {Item Name} to contain at most 1 word	Length Constraints.	"Ceramic Golden Swan/Elephant Vase Dry Flower Holder Arrangement Dining Table Home Decoration Accessories, Left Elephant"	Vase
4	Summarize {Item Name} to contain at most 4 words			Ceramic Golden Swan Vase
5	Summarize {Item Name} with Low specificity and to contain the words "Xbox Series S"	Phrase Inclusion Constraint.	"Skinit Decal Gaming Skin Compatible with Xbox Series S Controller - Officially Licensed NFL Dallas Cowboys Blast Design"	Xbox Series S Controller Skin
6	Summarize {Item Name} with Medium specificity and to contain the words "Compatible with Series S"			Skinit Decal Gaming Skin Compatible with Series S Controller
7	Summarize {Item Name} by dropping up to 10 words	Number of deleted words constraint.	"Girl Kayak Heartbeat Lifeline Monitor Decal Sticker 8.0 Inch BG 635"	Decal Sticker
8	Summarize {Item Name} with Medium specificity and by dropping up to 5 words			Girl Kayak Heartbeat Lifeline Monitor Decal Sticker

Table 1: Different instructions used by InstructPTS to generate product title summaries. Each instruction has different requirements that must be satisfied in the generated summary.

We follow a similar approach for generating product name summaries, and fine-tune FLAN-T5 models using instructions that are generated *automatically* from our parallel dataset of input product names and their corresponding summaries (cf. §3.2). Table 1 shows the instructions used for fine-tuning InstructPTS, as well as for generating product name summaries.

Using a product as a running example "Massage Orthopedic Puzzle Floor Mat for Kids Flat Feet Prevention Sea Theme 6 Elements", we describe in detail the instruction and the way they are constructed.

Specificity Level Constraints. Instructions 1–2 in Table 1 allow InstructPTS to generate summaries according to the specificity levels introduced in §3.2. These Low and Medium levels allow the model to dynamically determine the summary length based on the desired specificity. Depending on the original title, the Low specificity can yield summaries of slightly different lengths for different product. Our training data has different levels for the same input, which helps the model learn which words are important for each specificity.

Word Count. This instruction allows the model to generate summaries that contain up to a certain number of words. The instruction for training is constructed automatically, where for a product name and its ground-truth summary, depending on the number of words in the summary (k), we generate the instruction that has as a target the number of words equal to $k' = k + \Delta$ (Δ corresponds to a random integer $0 \leq \Delta \leq 3$, where $k > 3$). For instance, in the table below, the ground-truth summary contains 3 words, however, the instruction contains the constraint "at most 5 words". This allows the model to *flexibly* use 5 words or fewer as it sees fit, because sometimes the most coherent

summary may use fewer words due to the presence of multi-word phrases.

Summarize {Item Name} to contain at most **5** words. → Orthopedic Floor Mat

Instructions 3–4 in Table 1 show how the same name is summarized with 1 and 4 words. The choice of words is determined automatically by the InstructPTS model, allowing it to automatically pick the most salient words from the product name.

Phrase Inclusion. In real-world settings, depending on the context, certain words may be required in the summary (e.g. brand, size, color). We automatically construct instructions from the parallel dataset by randomly choosing a word or a sequence of words from the ground-truth summary. This allows InstructPTS to learn on how to incorporate specific phrases in the resulting summary. We evaluate the instruction following accuracy in §5.

Summarize {Item Name} with **Low** specificity and to contain the words "Orthopedic". → Orthopedic Mat

Instructions 5–6 in Table 1 show how the desired words are encoded in conjunction with categorical constraints. This allows the model to generate summaries of different specificity, and additionally enforce the inclusion of desired phrases.

Deletion of k -words. Instructions 7–8 in Table 1 allow deleting up to k -words. This represents the reverse case of the instructions that allow the model to output summaries of specific lengths. The instructions are inferred automatically from the ground-truth product name summary how many words need to be deleted, and additionally add a random integer $0 \leq \Delta \leq 3$.

Summarize {Item Name} by dropping up to **13** words. → Orthopedic Floor Mat

4 Experimental Setup

4.1 Evaluation Scenarios & Metrics

Automated Evaluation: For specificity constraints, we adopt BLEU and ROUGE metrics to automatically measure summary *quality* and their alignment with the ground truth. For other instructions, we compute the *instruction following accuracy* of InstructPTS, where we only assess if the model follows the constraints encoded in the instruction.¹ This verifies that the summary has the desired word count, or includes a specific phrase.

Human and Extrinsic Evaluation: We conduct human evaluation to assess summary quality (§6), and assess summary fidelity using retrieval (§7).

4.2 Baselines and Approach Setup

We compare InstructPTS against baselines that use different training strategies. We also assess different FLAN-T5 model sizes: (i) FLAN-T5-BASE, (ii) FLAN-T5-LARGE, and (iii) FLAN-T5-XL.

FLAN-T5-SFT: we perform supervised fine-tuning of FLAN-T5 models with input being the original product name, and the output being the ground-truth summary. This baseline is not controllable (e.g. specificity or number of words).

FLAN-T5-CC: We use Control Codes (CC) (Keskar et al., 2019) to guide summary generation. Each CC corresponds to a specific summarization instruction, enabling controllable summarization capabilities. We use the following CC: (i) `Low </s> {Item Name}`, and (ii) `Medium </s> {Item Name}`.

Training details: please see Appendix D for a detailed description of the training setup.

5 Automatic Evaluation Results

Table 2 shows the automated evaluation results on the 1M title test set. We compare different FLAN-T5 model sizes and the impact of the different training strategies. Output examples from InstructPTS are shown in Appendix A.

Text Generation Performance: A consistent pattern is that as model size increases, so do the BLEU and ROUGE metrics. For instance, FLAN-T5-XL improves by roughly 5 BLEU1 points over

¹We do not assess the accuracy of the instruction for deleting k -words, given that this task is designed to increase model robustness rather than downstream usage. Furthermore, determining the exact number of words to be deleted to generate valid summaries is not trivial and varies across product types.

FLAN-T5-BASE (for all strategies). We note a similar trend for ROUGEL.

Impact of Training Strategy: Training strategy has a significant impact. For the same model size, InstructPTS models obtain the best performance, e.g. InstructPTS with FLAN-T5-XL obtains an improvement of 13.3 BLEU1 points over the SFT and CC models. Finally, we note a convergence between CC and SFT for the FLAN-T5-XL models, with near identical performance. Our results show the advantages of instruction tuning for PTS.

Instruction Following: Table 3 shows the instruction following accuracy for different InstructPTS models, where we measure if the summary contains the desired number of words specified in the first instruction (I#1) or includes a specific phrase as specified in the second instruction (I#2) from Table 1. We find that the accuracy is significantly impacted by model size. FLAN-T5-XL obtains the highest instruction following accuracy among the FLAN-T5 models.

Summary Length: Table 4 shows the mean mean title length (number of words) and standard deviation for summarized titles generated for different summary types using InstructPTS (FLAN-T5-XL) on the entire test set. For specific word counts, we find that the model generally respects the maximum length imposed in the instruction. The categorical constraints have more variance compared to the specific word counts, and Medium summaries have an average length of 3.80 ± 1.28 words.

Compression Ratio: We also analyzed the data compression ratios for Low and Medium summaries based on character length. Results show high string compression ratios of 11:1 for Low and 5:1 for Medium summaries. We also observed that the compression ratio varies by product category, as shown in Appendix C.

6 Human Evaluation Study

To address the known limitations of automatic summarization evaluation, we perform a human study. We aim to answer the following questions:

- H1:** In a *pairwise comparison*, which model generates better product name summaries?
- H2:** Are the generated summaries *valid*?
- H3:** What is the *preferred* summary length by humans for a given product name?

Base Model	Strategy	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE1	ROUGE2	ROUGE3	ROUGE4	ROUGEL
FLAN-T5-BASE	SFT	0.455	0.309	0.180	0.115	0.571	0.358	0.161	0.074	0.570
	CC	0.451	0.307	0.176	0.114	0.567	0.356	0.156	0.073	0.566
	InstructPTS	0.585	0.411	0.247	0.160	0.665	0.450	0.230	0.118	0.663
FLAN-T5-LARGE	SFT	0.473	0.323	0.180	0.113	0.595	0.373	0.157	0.069	0.594
	CC	0.480	0.331	0.185	0.117	0.601	0.382	0.163	0.073	0.599
	InstructPTS	0.605	0.427	0.258	0.165	0.686	0.467	0.241	0.124	0.685
FLAN-T5-XL	SFT	0.509	0.356	0.196	0.120	0.634	0.408	0.173	0.075	0.632
	CC	0.509	0.357	0.195	0.120	0.633	0.408	0.172	0.075	0.632
	InstructPTS	0.642	0.463	0.277	0.173	0.718	0.502	0.258	0.127	0.716

Table 2: Text generation performance as measured based on BLEU and ROUGE metrics for the different training strategies and FLAN-T5 model sizes. In the case of CC and InstructPTS we can generate summaries according to the categorical constraints as in the ground truth (either **Low** or **Medium**), while for SFT we can only generate a single summary, which is compared against its ground-truth counterpart (either **Low** or **Medium**).

Model	Instruction	Acc
FLAN-T5-BASE	I#1 Summarize {Item Name} to contain at most k words.	0.674
	I#2 Summarize {Item Name} to contain the words "{T}".	0.618
FLAN-T5-LARGE	I#1 Summarize {Item Name} to contain at most k words.	0.673
	I#2 Summarize {Item Name} to contain the words "{T}".	0.714
FLAN-T5-XL	I#1 Summarize {Item Name} to contain at most k words.	0.765
	I#2 Summarize {Item Name} to contain the words "{T}".	0.760

Table 3: Instruction following accuracy for the different InstructPTS base models using instruction fine-tuning.

Summary Type	Summary Length
Low	2.07 \pm 0.76
Medium	3.80 \pm 1.28
1 Word	1.02 \pm 0.13
2 Words	1.95 \pm 0.36
3 Words	2.62 \pm 0.63
4 Words	3.06 \pm 0.94
5 Words	3.15 \pm 1.17

Table 4: The mean and standard deviation of the summarized title lengths (word count) for different summary types generated by InstructPTS (FLAN-T5-XL).

Data Evaluations are carried out on a sample of 10 popular product types (e.g. Electronics). For each product type we randomly sample 10 products and generate summary titles. Detailed evaluation setup is provided in Appendix B.

6.1 H1: Pairwise Summary Comparison

We compare the two best performing models, InstructPTS and CC using FLAN-T5-XL. For the same 100 product titles, we randomly generate either Low or Medium titles,² and ask the annotators to chose their preferred summary. To avoid position bias, the summaries are ordered randomly.

InstructPTS was preferred by the annotators in

²We compare only these two options, given that the FLAN-T5-XL-CC can only generate such summaries.

55% of the cases, while in 29% FLAN-T5-XL-CC model was preferred. In 12% the annotators chose *both* summaries being equally good, while in 4% of the cases, *neither* title was preferred. Finally, Cohen’s inter-rater agreement rate between two annotators was substantial with $\kappa = 0.61$.

6.2 H2: Validity of the Generated Summaries

Having established that InstructPTS generates the best summaries, two annotators judge if the summaries are valid. A summary is valid if it is *coherent* and can be used to *identify* at least the type of the original product.

We generate 7 different summary types per product. Table 5 shows the types and their validity scores. On this sample of 700 titles, Cohen’s inter-rater agreement was substantial ($\kappa = 0.69$).

Summary Type	Accuracy
Low	92.5%
Medium	97.5%
1 Word	39.5%
2 Words	78.0%
3 Words	85.0%
4 Words	90.0%
5 Words	96.0%

Table 5: Validity score (binary) of the different summary types for InstructPTS (FLAN-T5-XL).

The lowest scores are obtained by short summaries. The reason for that is that most products require two or more words for a summary to be meaningful w.r.t. the original product name, and be able to identify the original product. The highest scores are achieved for summaries of Medium specificity and those with 5 Words.

6.3 H3: Preferred Summary Length

In this study, we aim to better understand human preferences w.r.t. summary length for the different product categories. This can help determine the summary types InstructPTS should generate for different categories.

Table 6 shows the results in terms of length preferences by human annotators. We omit summaries that were deemed as not meaningful by the annotators (about 19%). The summaries are generated using the InstructPTS using FLAN-T5-XL model. We find a moderate agreement between annotators with a Cohen’s inter-rater agreement of $\kappa = 0.51$.

Across the different product categories, the preferences vary. For instance, for BEAUTY, the preferred summaries are longer, with 5 words. This is intuitive given the large variety of beauty products and brands. On the other hand, for FURNITURE, we see that an ideal summary length is with 2 words. Such products, in most cases, can be easily summarized with few words, e.g. “TV Stand”.

This study shows that ideal title summarization requires different lengths for different product categories. Our proposed InstructPTS model can robustly summarize products of any type using either Low or Medium summary specificity, which have variable summary length across product categories. Additionally, we can encode various constraints in terms of phrase inclusion in the summary. In 82% of cases Low summaries contain up to two words. Medium summaries on the other hand have more than three words in 78% of cases, with 57% having between 3 to 4 words. If we inspect the human preference of summary length in Table 6, we note that humans annotators tend to prefer summaries between 3–5 words, which represent summaries that have similar length as Medium summaries.

7 Extrinsic Evaluation with Retrieval

We have shown that InstructPTS can robustly summarize titles, following instructions for length and phrasal inclusion (cf. §3). To assess the fidelity of the summarized titles, we perform a retrieval-based extrinsic evaluation to determine how well the original products can be retrieved by using the summary titles. We hypothesize that a good summary with retain enough of the unique characteristics of the original product to be able to retrieve it. Additionally, this evaluation analyzes the trade-offs between summary length vs. ranking metrics of a target product under consideration.

Category	Preferred Length (Words)				
	1	2	3	4	5
BOOK	-	-	20%	-	80%
SHIRT	-	28.6%	28.6%	14.3%	28.6%
HOME	-	22%	22%	11%	44%
TOY FIGURE	-	37.5%	37.5	37.5%	-
SPORTING	-	-	62.5%	25%	12.5%
GOODS					
BEAUTY	-	25%	12.5%	25%	37.5%
TOOLS	12.5%	37.5%	50%	-	-
FURNITURE	-	100%	-	-	-
ELECTRONICS	-	33.3%	33.3%	33.3%	-
GROCERY	22%	67%	11%	-	-

Table 6: Summary preferences across product categories. Annotators pick their preferred summaries for a sample of 10 product names per product category.

Setup: We use a catalog of 5M products as our testbed. The product titles are summarized using InstructPTS (FLAN-T5-XL) with different instructions. The summary titles are then used as queries to review the top- k products in the catalog index using the BM25 algorithm. We also use the original title as an upper bound.

Evaluation: Evaluation is performed with standard IR metrics, Mean Reciprocal Rank (MRR) and Hit@ k . Higher values indicate that the summary retains more distinguishing information from the original product title.

Results: Table 7 shows the ranking scores of different summary types, based on a stratified sample of 100 products from over 800 different product categories (see Appendix C for more details). Intuitively, longer summaries obtain higher ranking scores than shorter summaries, since they tend to lose more information, leading to decreased ranking accuracy. Among all instructions, Medium achieves the best ranking scores. As shown in Table 4, Medium summaries are, on average, even longer than 5 Words summaries.

The MRR of 0.398 indicates that, on average, the ground-truth product is ranked in the 2nd and 3rd position. Furthermore, the Hit@20 score of 0.641 shows that in 64.1% of cases the ground-truth product is featured among the top 20 results. This study shows that our summaries retain key aspects that help identify the product in a set of 5M. It also provides guidance on how much the titles can be compressed.

Instruction	MRR	Hit@10	Hit@20
Original (upper bound)	0.991	0.998	0.999
Low	0.104	0.154	0.184
Medium	0.398	0.566	0.641
1 Word	0.008	0.010	0.016
2 Words	0.104	0.178	0.225
3 Words	0.220	0.345	0.416
4 Words	0.281	0.422	0.487
5 Words	0.286	0.416	0.480

Table 7: Ranking results for summaries generated by InstructPTS (FLAN-T5-XL). The first row is the upper bound, with the original product title used as a query.

8 Online Deployment

InstructPTS has been used in a leading global e-commerce service for various downstream shopping tasks. It can be applied for various content generation tasks related to product summarization, comparison, question suggestion, and review summarization. A 4k sample of generated content with embedded product titles from InstructPTS were evaluated for quality, and 96% were found to meet the validity criteria.

9 Conclusion

We presented InstructPTS, a new approach for Product Title Summarization, and demonstrated the effectiveness of instruction-tuning for this task. Through IFT we can train a highly accurate and controllable model for generating various types of summaries. Empirical studies using automatic and human evaluation studies showed that the model size has a significant impact in generating reliable and meaningful summaries, and at the same time it ensures the model’s ability to follow requirements specified in the instructions.

InstructPTS has been deployed in systems where product titles from a billion-scale catalog are summarized for various downstream applications, such as question answering and summarization. Future work will focus on more fine-grained instructions focusing on higher levels of specificity, and support for handling constraints based on brands/sizes/colors.

Limitations and Future Work

Our proposed approach has some limitations that we aim to address in future work. Namely, although the generated summaries are highly meaningful

and qualitative, they are constructed independently from their downstream applications. This creates a gap as to whether the most salient words for an application are chosen to be incorporated in a summary. For instance, for product retrievability, we aim at investigating whether choosing words to be incorporated in a summary can be provided by the BM25 ranking method, such that words with highest discriminative power are incorporated in the summary. We aim to do this in an end-to-end fashion, where the retrievability serves as a critic to the InstructPTS approach providing feedback on how to change the output summary.

Finally, we also aim to investigate the challenges in summarizing product names in conversational scenarios, where the requirements for product summaries change with every conversation turn.

References

- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *CoRR*, abs/2210.11416.
- Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.
- Yu Gong, Xusheng Luo, Kenny Q Zhu, Wenwu Ou, Zhao Li, and Lu Duan. 2019. Automatic generation of chinese short product titles for mobile display. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9460–9465.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Snehasish Mukherjee, Phaniram Sayapaneni, and Shankar Subramanya. 2020. Discriminative pre-

training for low resource title compression in conversational grocery. *arXiv preprint arXiv:2012.06943*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ohad Rozen, David Carmel, Avihai Mejer, Vitaly Mirkis, and Yftah Ziser. 2021. Answering product questions by utilizing questions from other contextually similar products. In *NAACL 2021*.

Fei Sun, Peng Jiang, Hanxiao Sun, Changhua Pei, Wenwu Ou, and Xiaobo Wang. 2018. Multi-source pointer network for product title summarization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 7–16.

Jingang Wang, Junfeng Tian, Long Qiu, Sheng Li, Jun Lang, Luo Si, and Man Lan. 2018. A multi-task learning approach for improving product title compression with user search log data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Manyi Wang, Tao Zhang, Qijin Chen, Chengfu Huo, and Weijun Ren. 2020. Selling products by machine: a user-sensitive adversarial training method for short title generation in mobile e-commerce. *DLP-KDD*, page 9.

Joan Xiao and Robert Munro. 2019. Text summarization of product titles. In *eCOM@ SIGIR*.

Jian-Guo Zhang, Pengcheng Zou, Zhao Li, Yao Wan, Xiuming Pan, Yu Gong, and Philip S Yu. 2019. Multi-modal generative adversarial network for short product title generation in mobile e-commerce. *arXiv preprint arXiv:1904.01735*.

Appendix

A Example InstructPTS Summaries

Table 8 shows example summaries generated by the InstructPTS model using FLAN-T5-XL as a base model. For each product name, 7 different summary types are generated.

Product Title	Summary Type	Generated Summary
New Balance Men's Fresh Foam Arishi V3 Classic Running Shoe, Black/Wave, 8.5	1 Word	Shoe
	2 Words	Running Shoe
	3 Words	New Balance Shoe
	4 Words	New Balance Running Shoe
	5 Words	New Balance Men's Running Shoe
	Low	Running Shoe
	Medium	New Balance Running Shoe
Happy Belly Frozen Chopped Kale, 12 Ounce	1 Word	Kale
	2 Words	Chopped Kale
	3 Words	Happy Belly Kale
	4 Words	Happy Belly Frozen Kale
	5 Words	Happy Belly Kale
	Low	Kale
	Medium	Happy Belly Kale
Vinyl Wall Art Decal - Thankful - 10.5" x 23.5" - Trendy Autumn Harvest Fall Leaves Seasonal Quote for Home Bedroom Kitchen Dining Room Office Church Decoration Sticker (Orange)	1 Word	Decal
	2 Words	Wall Art
	3 Words	Wall Art Decal
	4 Words	Vinyl Wall Art Decal
	5 Words	Vinyl Wall Art Decal
	Low	Wall Art Decal
	Medium	Vinyl Wall Art Decal
Honbay 4PCS 18mm Rubber Replacement Watch Band Strap Loops (Black)	1 Word	Watch
	2 Words	Watch Band
	3 Words	Honbay Watch Band
	4 Words	Watch Band Strap Loops
	5 Words	Watch Band Strap Loops
	Low	Watch Band Strap
	Medium	Honbay Watch Band Strap
DECOHS 2 Packs Hanging Flower Basket Frost Cover-27.5 x 39 Inch Large Dual Drawstring Plant Protection Cover Bags-Hanging Plant Pots Frost Cover Protecting Plants from Freezing Animals Eating	1 Word	Frost
	2 Words	Flower Basket
	3 Words	DECOHS Flower Basket
	4 Words	DECOHS Hanging Flower Basket
	5 Words	DECOHS Flower Basket Frost Cover
	Low	Frost Cover
	Medium	DECOHS Flower Basket Frost Cover
Mens Retired Baseball Coach Shirt. Free to Do Whatever Retirement T-Shirt	1 Word	T-Shirt
	2 Words	Coach Shirt
	3 Words	Baseball Coach Shirt
	4 Words	Retired Baseball Coach Shirt
	5 Words	Retired Baseball Coach Shirt
	Low	T-Shirt
	Medium	Retired Baseball Coach Shirt
ELISORLI Compatible with Xiaomi Redmi Note 11 Pro 4G/5G Wallet Case Leather Wrist Strap Lanyard Flip Cover Card Holder Stand Phone Cases for Redme Note11 11E 11Pro Cell Accessories Women Men Black	1 Word	Phone
	2 Words	Phone case
	3 Words	ELISORLI Phone Case
	4 Words	ELISORLI Compatible with Xiaomi
	5 Words	ELISORLI Phone Case
	Low	Phone case
	Medium	ELISORLI Phone Case
Olive Loves Apple Promoted to Big Sister Colorful Announcement T-Shirt for Baby and Toddler Girls Sibling Outfits Chill Shirt	1 Word	T-Shirt
	2 Words	Olive T-Shirt
	3 Words	Olive Loves Apple
	4 Words	Olive Loves Apple T-Shirt
	5 Words	Olive Loves Apple Promoted
	Low	T-Shirt
	Medium	Olive Loves Apple Promoted to Big Sister

Table 8: Example summaries generated by the InstructPTS model. For each product name we show 7 different summary types that are generated.

B Human Evaluation Setup

In §5 we showed the results from three human evaluation studies. The studies captured the intrinsic quality of summaries. In H1, we compared the two best performing models to determine which summaries were preferred by human annotators. While in H2 and H3, for the best performing model, we captured validity and summary length preference by annotators.

Here we describe in detail the human evaluation setup. We carry out the annotation using two expert human annotators. In the human evaluation studies, we focus on 10 popular e-commerce product types such as: BOOK, SHIRT, HOME, TOY FIGURE, SPORTING GOOD, BEAUTY, TOOLS, FURNITURE, ELECTRONICS, and GROCERY.

H1: Pairwise Summary Comparison

For the two best performing models, InstructPTS (FLAN-T5-XL) and FLAN-T5-XL-CC, and the summary types Low and Medium, we compare which outputs are preferred by annotators.

For the sample of 10 product categories, we sample randomly 10 products, and for each of the product names generate their corresponding Low and Medium summaries for the two models under comparison. We randomly pick either the Low or Medium summary from both models for the same product for comparison. This results in a total of 100 annotations by two expert annotators.

To avoid any potential position bias, we shuffle the order in which the summaries are shown the annotators, and the model information, which produces the summaries is kept hidden from the human annotators.

An example preview of the annotation job is shown in the Table 9 below:

Product Name	Summary A	Summary B	Label
"BushKlawz Premium Prince Beard Oils Variety Set Pack Bundle of Full Size 2 oz Lumber Pacific and Urban Prince Scents and Naked Prince Scent Fragrance Set Bundle Kit"	BushKlawz Beard Oils	Premium Beard Oils	- Summary A - Summary B - Both - Neither

Table 9: Annotators in this pairwise comparison choose their preferred summary, without being aware of the model that produced it. In this case summary A is generated by InstructPTS (FLAN-T5-XL), while summary B is produced by FLAN-T5-XL-CC.

H2: Validity of the Generated Summaries?

In this study, we asked the human annotators to judge whether a summary is meaningful. We defined meaningfulness as a summary which is *coherent*, it can be used to *identify* the product or the product type/family.

We analyzed only the summaries generated by InstructPTS with FLAN-T5-XL as established through automated metrics, as well as the human evaluation in H1. We asked two human annotators to judge the meaningfulness of the summaries for 100 products (10 random products from 10 product categories), which resulted in a total of 700 summaries (each product name is summarized using 7 different summary types).

To judge the meaningfulness score, the annotators are shown the summary along with the original product name for judgement. Table 10 shows an example of the annotation task.

Product Title	Type	Summary	Is Meaningful?
"Fresh Products Bio Conqueror 105 Enzymatic Odor Counteractant Concentrate FRS 12-32BWB-MG"	Low	Odor Counteractant Concentrate	Yes No
	Medium	Fresh Products Odor Counteractant Concentrate	Yes No
1 Word	Odor		Yes No
2 Words	Odor Counteractant		Yes No
3 Words	Fresh Products Odor		Yes No
4 Words	Odor Counteractant Concentrate		Yes No
5 Words	Fresh Enzymatic Odor Counteractant Concentrate		Yes No

Table 10: Annotators judge for each summary type for the given product, if the resulting summary is meaningful.

H3: Preferred Summary Length

In this study, we gather the preference of human annotators in terms of summary length. Here too as in the previous studies, we sample 10 products from 10 product categories, and ask two human annotators to provide their preferred summary for a given product, among the 7 different summary types. Here too, the study only analyzes the summaries generated by InstructPTS with FLAN-T5-XL, given that only this model can support the flexible generation of different summary types. Example of the annotation task is shown in Table 11.

Product Name	Summaries	Preferred Summary
	Sneaker	1 Word
Adidas Ultraboost	Adidas Sneaker	2 Words
6.0 DNA X Parley	Adidas Running Shoe	3 Words
Non-Dyed/Non-Dyed/Non-Dyed	Adidas Ultraboost DNA X	4 Words
8.5 D (M)	Adidas Ultraboost DNA X Parley	5 Words
	Running Shoe	Low
	Adidas Ultraboost DNA X Parley	Medium

Table 11: Annotators provide their preferred summary type for a given product name, shown in the order {Low, Medium, 1 Word, 2 Words, 3 Words, 4 Words, 5 Words}.

C Retrieval Results by Product Category

For extrinsic evaluation (§7), we utilized a real e-commerce product catalog, indexing a total of 5M products. To ensure an unbiased evaluation of the retrieval results presented in Table 7, we took a stratified sampling approach where 100 products were randomly selected from each product category. This method helped mitigate any potential biases caused by variations in the popularity of different product categories.

We selected 25 product categories and show their product-level MRR scores by InstructPTS (FLAN-T5-XL) in Table 12, ranked by the relative decrease of MRR when transitioning from Medium to Low specificity:

$$\frac{MRR(\text{Medium}) - MRR(\text{Low})}{MRR(\text{Medium})} \quad (1)$$

Additionally, to understand how much product titles are compressed, we calculate the data compression ratio (CR) of the original titles using:

$$CR = \frac{\text{len}(\text{original product title})}{\text{len}(\text{summarized title})} \quad (2)$$

where the $\text{len}()$ function is the string length of the titles in characters.

The results show significant variations in CRs and MRR scores across different product categories. Notably, product categories such as BEAUTY and GROCERY exhibit relatively lower CRs and the difference of CRs between Low and Medium is smaller compared to other product categories. This phenomenon can be attributed to the fact that the ground-truth of Low summaries does not further delete more words compared with Medium, since excessively deleting words from their names may render them less identifiable. Therefore, the ranking scores are relatively higher, compared to product categories like EARRING and SHIRT, whose CRs of Low specificity can be up to 18.

D Training Details

All models are trained for a maximum of 50 epochs, with an early stopping criterion of 5 epochs of non-decreasing loss on the validation set. The batch size was set to 32.

We used AdamW (Loshchilov and Hutter, 2017) to optimize the model’s parameters. The learning rate was set to $lr = 2e^{-4}$, with a 10% of steps from the first epoch used as a linear warm-up stage to find the optimal starting lr .

Product Category	MRR (Low)	MRR (Medium)	CR (Low)	CR (Medium)
SHIRT	0.000	0.280	12.841	4.651
EARRING	0.001	0.288	16.021	6.620
NECKLACE	0.002	0.322	14.725	6.276
CELLULAR PHONE	0.025	0.318	15.849	5.834
RING	0.020	0.234	18.025	6.199
FURNITURE	0.039	0.451	12.193	6.178
MASSAGER	0.052	0.550	11.478	6.246
TEA	0.104	0.735	11.813	4.625
CANDLE	0.059	0.393	14.571	5.537
WRENCH	0.093	0.544	6.932	3.333
SPEAKERS	0.091	0.524	8.603	4.891
PAINT	0.060	0.308	8.770	3.797
DRIED PLANT	0.097	0.470	11.355	6.642
HAIR EXTENSION	0.067	0.306	10.827	6.106
TOY FIGURE	0.105	0.524	10.808	4.786
GUITARS	0.094	0.416	8.193	4.966
TOOLS	0.124	0.506	8.089	4.298
CONSUMER ELECTRONICS	0.124	0.503	8.577	5.010
PRINTER	0.102	0.396	9.980	5.536
SPORTING GOODS	0.120	0.446	8.835	4.082
HOME	0.150	0.486	11.160	5.134
MEAT	0.264	0.832	7.588	2.945
FRUIT	0.268	0.834	9.297	3.883
BEAUTY	0.202	0.540	9.288	5.170
GROCERY	0.299	0.767	6.890	3.079

Table 12: MRR scores and compression ratios (CR) for different product categories. The order of product categories is determined by Eq. 1 in descending order.