# Cross-Attribute Consistency Detection in E-commerce Catalog with Large Language Models

**Anonymous ACL submission**

## Abstract

Comprehending the quality of data represented on an E-commerce product page is a challenge and is currently achieved with varied approaches that are dependent on large task-specific datasets curated with human efforts. This slows down the process of scaling to a large catalog scope. The recent advancements in Large Language Models (LLM) have revolutionized their ability to significantly enhance various downstream applications using small and carefully curated datasets. In this paper, our focus is to explore LLM capability in addressing a challenge related to the catalog quality assessment. To be specific, we aim to detect the consistency of information presented between Unstructured Attributes (UA) (incl. Title, Bullet Points (BP), Product Description (PD)), and Structured Attributes (SA) within a product page through pairwise evaluations using predefined class labels. To achieve it, we propose a novel approach, CENSOR, that utilizes LLM in two phases. In the first phase, off-the-shelf LLM is leveraged in a zero-shot manner using prompt engineering techniques. While in the second phase, open-source LLM is fine-tuned with a small human curated dataset along with the weak labeled data generated in first phase as a data augmentation technique to incorporate domain-specific knowledge. The fine-tuned LLM overcomes the deficiencies observed in the first phase and entails the model to address the consistency detection task. Evaluation conducted using the E-commerce dataset which include a comprehensive set of 186 distinct combinations of <Product Type, SA>, CENSOR fine-tuned model outperforms the baseline method and CENSOR zero-shot model with +34.4 and +19.4 points on F1-score respectively.

## 1 Introduction

An E-commerce catalog contains several products which are described using a set of attributes. In general, attributes can be broadly divided into two different types mainly Unstructured Attributes (UA) and Structured Attributes (SA). UA provides information using unstructured data such as product text (Title, Bullet Points, and Description), images, and videos. While the goal of SA is to provide a summary of product information useful for other tasks such as product search, discovery, and advertising. However, due to many reasons, consistency may lack between the information mentioned in UA about the SA. Figure 1 showcases a scenario where **material** value mentioned in Textual UA (Title) providing contradicting information with SA.

Such Cross-attribute consistency not observed between UA and SA can have different challenges for the end consumer such as: (1) Confusion and impact their buying decision and (2) Increases the returns of the sold products due to mismatch in expectations. Addressing the aforementioned and other similar cases will provide significant cost and time benefits.

Earlier research (More, 2016; Maadan et al., 2016; Zheng et al., 2018; Xu et al., 2019; Mehta et al., 2021) targeted comprehending SA information present in a textual UA by extracting SA information from UA. The main purpose of these works is to fill the missing information in the catalog to achieve high completeness. However, they do not focus on verifying the consistency of the SA information already provided by the seller with the rest of the product UAs. This is the major focus of our research and aim to address the issue at scale.

Although aforementioned research is proven effective for completing missing SA information. They still poses challenges in extracting accurate information from UA due to confusing attributes and diverse surface forms. Hence, extracting SA from product UAs and then comparing it with the seller provided SA might not be an effective approach for detecting inconsistency across Cross-level attributes. Therefore, in this work, we target an end-to-end solution for detecting inconsistency i.e.,

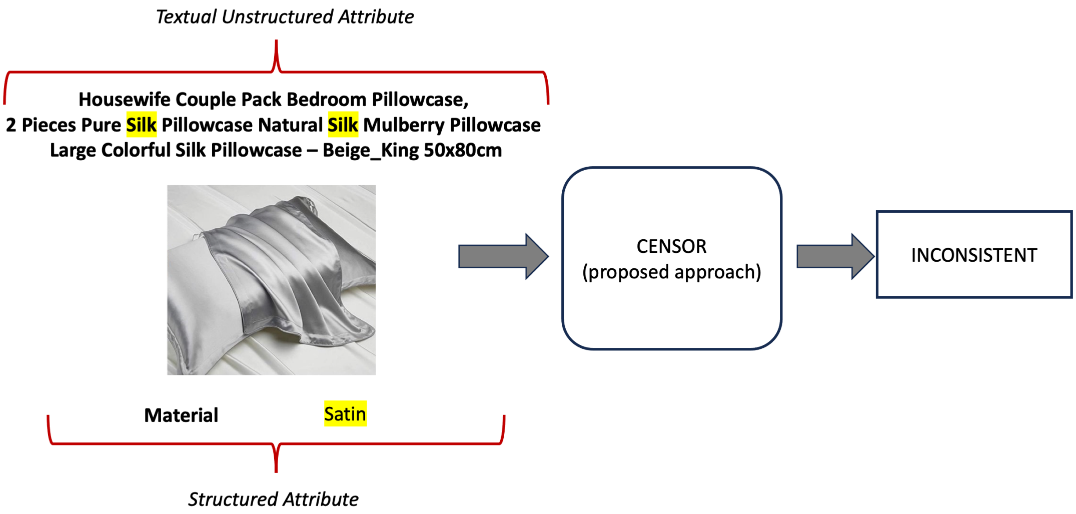
Textual Unstructured Attribute
Housewife Couple Pack Bedroom Pillowcase, 2 Pieces Pure Silk Pillowcase Natural Silk Mulberry Pillowcase Large Colorful Silk Pillowcase – Beige_King 50x80cm
CENSOR (proposed approach)
INCONSISTENT
Material
Satin
Structured Attribute

Figure 1: Example showcases the Inconsistency detected by proposed approach (CENSOR, details in Section 3) between Textual UA (Title) mentioning **material** value as *Silk* and separately mentioned in SA as *Satin*.

comparing UA usually present as a larger sequence of tokens with SA constituting a smaller sequence of tokens with a **C**ross-attribut**E** Co**NS**istency Detect**OR** (henceforth, CENSOR).

Our proposed approach is divided into two different phases. In the first phase, we build CENSOR-zero-shot with an off-the-shelf Large Language Models (LLM) (Anthropic, 2023; OpenAI, 2023) and prompting techniques to produce predefined class labels. This phase provide us with a baseline approach for the cross-attribute consistency detection task and also help to generate synthetic data for the second phase. While in the second phase, we leverage open-source LLM (Chung et al., 2022; Penedo et al., 2023; Touvron et al., 2023) as a generative formulation to produce CENSOR-fine-tuned model by fine-tuning on the human curated labeled dataset. To be specific, CENSOR-fine-tuned is built as a generative model which takes <UA, SA> pair and other product relevant information in a sequence as an input and returns a predefined class label. The CENSOR-fine-tuned is robust in handling diverse SA's which include different surface forms, measurement units, and varied values.

The main contributions of this paper are:

- We propose CENSOR, a two-phase approach to identify the consistency across SA and a textual UA.

- We introduce different prompting strategies for cross-attribute consistency detection task with CENSOR-zero-shot.

- We present experimental results on an E-commerce dataset with diverse SA to showcase the efficacy of CENSOR-fine-tuned.

To the best of our knowledge, this is a first attempt at building an approach to detect consistency between a SA and textual UA with a generative formulation for a large-scale E-commerce catalog. The rest of the paper is organized as follows. We describe related work which closely aligns with our work in Section 2. Further, in Section 3 we present our proposed method and its variants. We describe the experimental setup in Section 4 and discuss our findings in the Section 5. Finally, we conclude our observations in Section 6.

## 2 Related Work

In the related work, we mention those works which are closely related to our task.

### 2.1 E-commerce Attribute Extraction

Several works have been proposed earlier (Kannan et al., 2011) to extract the SA information from the product title and description. Most of the works proposed the E-commerce Attribute Extraction problem as a special case of Named Entity Recognition (NER) task (More, 2016; Wang et al., 2020). Zheng et al. (2018) proposed an OpenTag architecture based on the combination of Bi-LSTM and Conditional Random Fields (CRF) for extracting a different set of attributes and not targeted specifically for numeric attributes. Similarly, Xu et al. (2019) focused on scaling up extraction and empowered attribute value extraction from product title using an attribute-comprehension-based approach. Multimodal extraction is also proposed (Zhu et al., 2020) to complement different modalities such as product images and descriptions for extraction of attributes. Nevertheless, the approach is not specific for numeric attributes and mostly concentrated on those attributes where the useful visual information from product images has an impact. However, for extracting only numeric attributes from the product description, Mehta et al. (2021) designed a platform by leveraging distant supervision. As discussed, extracting numeric SA for inconsistency detection is ineffective, we would be comparing cross-level numeric SA and textual UA directly and classifying them into predefined classes.

### 2.2 E-commerce Text Classification

Many works are proposed for labeling certain text with predefined labels (Sun et al., 2019). However, we want to highlight those classification works which specifically leverage the e-commerce text that spans different levels. Tan et al. (2020) utilized product title and descriptions for product categorization into leaf category using a machine translation-based approach. Other approaches (Zhao, 2020) used customer reviews present in languages other than English and performed sentiment analysis. In this paper, we portray inconsistency detection of cross-level attributes into a classification problem setup.

### 2.3 Large Language Models for E-commerce

At present, LLM are becoming common for understanding and generating human language. They are built using transformer architecture (Vaswani et al., 2017) using different variations such as Sequence-to-Sequence (a.k.a Encoder-Decoder) (Chung et al., 2022) and Decoder-only (Zhang et al., 2022; OpenAI, 2023; Touvron et al., 2023). Recently, focus on applying LLM for E-commerce specific tasks has increased. Maragheh et al. (Maragheh et al., 2023) used LLM as augmenter for recommendation-related tasks. Similarly, LLM is used for relationship identification in E-commerce specific knowledge graph completion models (Chen et al., 2023). There are works (Li et al., 2023) which expanded LLM with instruction-tuning targeted for E-commerce data for several downstream tasks such as Named Entity Recognition (NER), Review Topic Classification and so on. Our work also leverages LLM for E-commerce domain, however, our focus is to comprehend the quality of E-commerce data by identifying Cross-attribute consistency of Cross-level information (Jurgens et al., 2014).

## 3 Approach

In this section, we present CENSOR variants whose aim is to generate predefined labels by detecting consistency observed across <UA,SA> pairs.

### 3.1 CENSOR - Problem Formulation

Let $\mathcal{S}$ represent a set of SA, and $\mathcal{U}$ an UAs . Each product $p \in \mathcal{P}$ can be thought-of as a textual representation of a product comprising relevant information about the product; e.g., Product Type it belongs to. We set forth the CENSOR as "Text-to-Text" generation inspired by previous unifying frameworks for Natural Language Processing (NLP) tasks (Raffel et al., 2020; McCann et al., 2018) and their effectiveness demonstrated for the classification task. Given any product-related representation $p \in \mathcal{P}$ containing Structured Attribute $s \in \mathcal{S}$, Unstructured Attribute $u \in \mathcal{U}$ and a target output $c \in \mathcal{C}$, learn a function:

$$f : \mathcal{P} \times \mathcal{U} \times \mathcal{S} \rightarrow \mathcal{C} \tag{1}$$

Unlike other architectures, which typically require training a task-specific layer (e.g., classification (Nogueira et al., 2020)) from scratch beyond backbone model, "Text-to-Text" formulation can leverage the pre-trained network's capacity for generating output tokens based on pretrained knowledge, saving time and resources. Therefore, utilizes LLM with two different ways. Figure 2 presents the overall framework.

### 3.2 CENSOR - Zero-shot

#### 3.2.1 Methodology

CENSOR-Zero-shot utilize off-the-shelf LLM in a zero-shot manner for Cross-attribute Consistency Detection task with prompt engineering techniques. Prompt is the function (Equation 1) we design to address the task. CENSOR-Zero-shot is built on a hypothesis that off-the-shelf LLM have been pretrained on vast amounts of textual data, and pose a rich contextual understanding. Leveraging such contextual knowledge will be advantageous in mitigating ambiguity observed for the propsed task.

CENSOR-Zero-shot is built with a two-step approach, mainly *Prompt Construction* and *Handling*

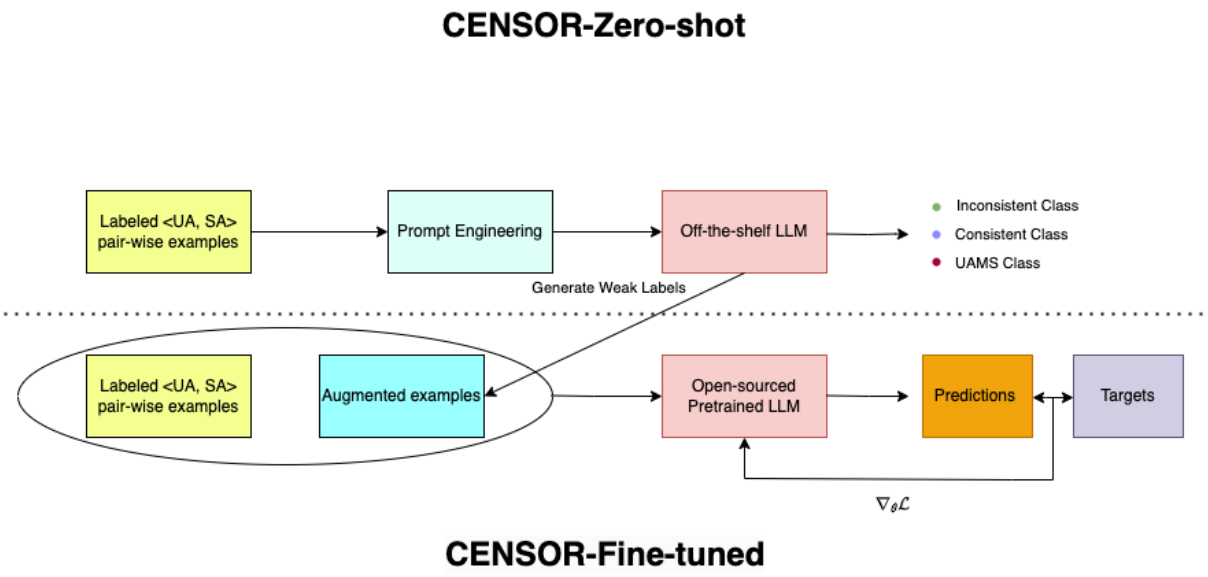

Figure 2: CENSOR Framework.

*Non-Deterministic Behavior*. We discuss each of them in the following sections.

#### 3.2.2 Prompt Construction

For the prompt construction, we employed Chain-of-Thought (CoT) (Wei et al., 2022) prompting technique. In our CoT driven prompt, an important intermediate reasoning step involves enabling an off-the-shelf LLM to understand the relationship between UA and SA. Hypothesis here is that if the LLM fails to identify a relationship, then its prediction is prone to be a hallucination. Additionally, this intermediate reasoning step provides an opportunity that let's LLM predict "I don't know" to prevent any hallucinations.

When the LLM identifies that UA and SA are relevant, the prompt goes on to categorize it into `CONSISTENT` predefined class as UA might contain SA or expresses a similar meaning as SA. This emphasis on the "containing" relationship stems from our observation that off-the-shelf LLMs excel at detecting consistency in these readily distinguishable cases, even when dealing with typically lengthy UAs (Product Description). If UA and SA are relevant but do not share similar meanings, nor does UA contain SA, the prompt would classify this pair as `INCONSISTENT` predefined class. For the undecided cases or not relevant cases, the prompt classifies them under the `UAMS` predefined class (more details about predefined classes is presented in Section 4.1).

To gain deeper insights into the reasoning behind predefined class predictions made by the off-the-shelf LLM in a zero-shot manner, we also attain prediction justification as possible reasons from the LLM.

#### 3.2.3 Handling Non-deterministic Behavior

Earlier research found that sometimes off-the-shelf LLM could be non-deterministic, even when the context, instructions, input data remain the same in the prompt. To address it, we set the temperature hyper-parameter as 0 and execute the prompt 3 times for deciding the predefined label using majority voting.

### 3.3 CENSOR - Fine-tuned

We observed that the performance of `CENSOR-Zero-shot` is limited as it cannot effectively comprehend the domain of E-commerce catalog. Hence, there is a requirement for a solution that is adapted for E-commerce domain and also understand the Cross-attribute Consistency Detection task in an efficacious manner. Therefore, we designed `CENSOR-Fine-tuned` to incorporate domain-specific E-commerce knowledge and also learn about the Cross-attribute Consistency Detection task by using both human curated gold standard data along with the synthetically generated weak labeled data.

#### 3.3.1 Model Architecture

Following Equation 1, we design the `CENSOR-Fine-tuned` based on the Encoder-Decoder architecture (Chung et al., 2022). Recent studies (Fu et al., 2023) have shown that Encoder-Decoder architecture outperforms Decoder-only architecture (Deng et al., 2023) due to attention degeneration issue. Also, Encoder-Decoder architecture help us to explore the potential of incorporating reasons generated by off-the-shelf LLM to add more context for an <UA,SA> pair along with the Product Type.

#### 3.3.2 Decoder Variants

Motivated by earlier works (Puri and Catanzaro, 2019; Nogueira et al., 2020) demonstrating that the choice of Decoder tokens in an Encoder-Decoder architecture can have a significant influence on generation outcomes, especially in data scarcity setting. CENSOR-Fine-tuned explores three different Decoder options based on three different hypothesis.

**Label-only** Our first hypothesis use the Decoder (output as in Equation 1) tokens as predefined Class Label. To avoid any pretrained knowledge interference with the prediction, we further modify the Decoder tokens to make them as unique (or special tokens).

**Label+Template_Reason** To add additional context to the Decoder along with predefined Class Label, we propose second hypothesis. In this approach, context is identified using the segment in an UA that is semantically closer to the SA using Sentence Embeddings (Reimers and Gurevych, 2019). The Decoder is then formulated as a "Label+Template_Reason", which includes both context i.e., segment in an UA and the predefined Class Label.

The intuition here is that if relevant UA and SA are consistent (referred to ground-truth class), then based on relevant UA and product context, the generated content should be semantically close to SA.

**Label+LLM_Reason** Our third hypothesis attempts to utilize the reason generated by off-the-shelf LLM when predicting predefined labels. The intuition here is that the reasons provided by an off-the-shelf LLM can be seen as an additional context which can benefit the Cross-attribute Consistency Detection task.

We observed with CENSOR-Zero-shot that the reasons it generates provide information about relational knowledge between UA and SA. Hence, we aim to include those reasons for each training data point. Therefore, Decoder of CENSOR-Fine-tuned is formulated as a "Label+LLM_Reason", which includes both LLM reason and the predefined class label.

## 4 Experimental Setup

In the following, we present the dataset information and the evaluation metrics used for comparison.

### 4.1 Predefined Class Labels

Three different class labels are used. If the SA value is consistent with the UA, the CONSISTENT label is used. While if the SA value is inconsistent with the UA, the INCONSISTENT label is used. If the SA value is missing from the UA, the UAMS label is used, which indicates that UA doesn't have any presence of SA. For example, color SA is never mentioned in the Title of the product, it is considered the predefined label of <color, Title> pair is UAMS.

### 4.2 Dataset

We targeted 186 distinct combination of <Product Type, SA> from a English locale country that cover variety of products[1]. We collected the dataset by leveraging our organization's annotators by providing them with an <UA, SA> pair to annotate predefined labels. As described in Section 4.1 , CONSISTENT, INCONSISTENT, and UAMS are the predefined labels used for annotating an <UA, SA> pair.

The entire dataset is split into 66% for training and 34% for the testing. Further, the training set is split into 65% for training and 35% for the validation.

We also collected synthetic data using an off-the-shelf LLM with prompt engineering techniques presented in Section 3.2.2 for expanding the training data. This is motivated by following reasons:

- Combining augmented data with the actual training data allows for training a more compact and computationally efficient model.
- Overcome the human curated data collection challenges such as scalability, bias, and cost.

Due to augmentation with the synthetic data the training data overall size has increased by 68.5%. We have kept the same size for the validation and testing set to clearly identify the benefit from the synthetic labels.

[1] Randomly sampled according to the task requirement and do not reflect the overall quality.

### 4.3 Evaluation Metrics

We evaluate the performance of the proposed approaches with the weighted averages of Precision, Recall and F-1 score. With weighted averaging, the output measurements, (i.e., Precision, Recall, or F-1 score), have accounted for the contribution of each class as weighted by the number of examples of that given class.

- **Weighted Precision:** Weighted mean of precision with weights equal to class probability.
- **Weighted Recall:** Weighted mean of recall with weights equal to class probability.
- **Weighted F-1 Score:** Weighted mean of F1-measure with weights equal to class probability.

## 5 Experimental Results

### 5.1 Methods

We design a simple Encoder-only model baseline termed as Expandable Validation (EVA). It is built in a two-step process. In the first step, self-supervised learning (SSL) (Gui et al., 2023) is leveraged to adapt the pretrained Encoder-only ALBERT Base v2 (Lan et al., 2020) model to E-commerce domain without any manually curated labels. While in the second step, the domain adapted model is fine-tuned with the actual dataset.

We consider the CENSOR-Zero-shot which utilizes Claude-v1.3 (Anthropic, 2023) as the off-the-shelf LLM with a given prompt (Section 5.3) as the strong baseline. While, CENSOR-Fine-tuned which leverages Flan-T5-XL (Chung et al., 2022) as its Encoder-Decoder architecture as the improvements proposed over it.

### 5.2 CENSOR-Fine-tuned Implementation

We fine-tuned CENSOR-Fine-tuned variants for 3 epochs with a constant learning rate of $10^{-4}$ using AdamW optimizer. Decoder length was varied between 10 and 512 tokens based on the Decoder variants. We used BLEU score to comprehend the fine-tuned check-point model quality at the end of each epoch during fine-tuning and selected the best saved check-point for inference. All Experiments were executed on 8 NVIDIA A10G GPUs (each 24GB). To conserve memory and accelerate the fine-tuning process, all models were fine-tuned using the BF16 format.

### 5.3 CENSOR-Zero-shot Prompt

In the following, we present the Chain-of-Thought (CoT) (Wei et al., 2022) prompt constructed for prompting Claude-v1.3[2].

- **Step-1**: Extract values in column 'UA' {UA Value} and column 'SA' {SA Value}. Do not return anything for this step.
- **Step-2**: Use your knowledge to understand the relationship between value {UA Name} and {SA Name}. Do not return anything for this step.
- **Step-3**: Determine whether {UA Value} includes or expresses a similar meaning as {SA Value}, relying on your knowledge and considering the relationship you identified between value {UA Value} and {SA Value}. If you cannot find a relationship between {UA Value} and {SA Value} from Step-2, your prediction answer should be 'UAMS', If (you think the meanings of {UA Value} and {SA Value} are relevant and similar) or (if value {UA Value} contains value {SA Value}), your prediction answer should be 'Consistent', If (you are very confident that the meanings of {UA Value} and {SA Value} are relevant, but (the meanings of {UA Value} and {SA Value} are not similar)) and (value {UA Value} does not contain value {SA Value}), your prediction answer should be 'Inconsistent' Otherwise, your prediction answer should be 'UAMS'. If you cannot decide, your prediction answer also should only be 'UAMS'. Do not return anything except 'Consistent', 'Inconsistent' or 'UAMS'.
- **Step-4**: Collect the reason for your prediction. Let's work this out in a step by step way to be sure we have the right prediction answer
- **Step-5**: Do not include any information generated from the above steps in the output.
- **Step-6**: After you provide the reason, provide all your response in JSON format with the following keys: 'reason', 'prediction', 'product_id' from column 'product_info'({product_info}). Do not return anything except JSON.

[2] At the time of experimentation, Claude-v2.0 was not available.

### 5.4 Results and Discussion

Overall results attained using different methods are presented in the Table 1 and Table 2[3].

Table 1: Overall Prediction Results (Weighted Precision and Weighted Recall) showing improvements over Encoder-only baseline

| Model | Precision (↑)<br>(Weighted) | Recall (↑)<br>(Weighted) |
|---|---|---|
| CENSOR-Zero-shot<br>(Claude-v1.3) | +21.9 | +6.4 |
| CENSOR-Fine-tuned | | |
| +Label | **+46.1** | +16.6 |
| +Label+Template_Reason | +32.4 | +7.8 |
| +Label+LLM_Reason | +41.6 | +5.5 |
| CENSOR-Fine-tuned<br>+Synthetic Data+Label | +44.3 | **+22.4** |

Table 2: Overall Prediction Results (Weighted F1-score) showing improvements over Encoder-only baseline

| Model | F1-score (↑)<br>(Weighted) |
|---|---|
| CENSOR-Zero-shot<br>(Claude-v1.3) | +15 |
| CENSOR-Fine-tuned | |
| +Label | +33.8 |
| +Label+Template_Reason | +19 |
| +Label+LLM_Reason | +25.4 |
| CENSOR-Fine-tuned<br>+Synthetic Data+Label | **+34.4** |

We found that CENSOR-Zero-shot can directly compare the relevant UA and SA values despite their length discrepancy. Furthermore, it can provide reasons for classification that are human interpretable with a few or even no post-processing. Such effective characteristics makes it a suitable stand-alone approach.

Also, we present the UA-wise results of proposed approaches in the Table 3 and per class results in the Table 4.

Effectiveness of data augmentation through the inclusion of synthetic data generated by off-the-shelf LLM has been demonstrated in the Table 1 and Table 2 . CENSOR-Fine-tuned with data augmentation outperforms CENSOR-Zero-shot, and the same method without data augmentation on weighted avg. F-1 score. This can be accredited to the quality of synthetic data created by handling non-deterministic behavior of off-the-shelf LLM (discussed in Section 3.2.3) that retained consistent responses as augmented instances.

CENSOR's advantages shine through in its ability to substantially cut down human annotation costs and the time needed for crafting top-notch synthetic training data. Specifically, compared to the variant with no synthetic data augmentation, adding more synthetic data improves the weighted avg. Recall by a significant margin, which indicates the synthetic data generation can effectively capture more relevant instances. The high recall is especially meaningful in the E-commerce domain since the goal of identifying consistent information is crucial to reduce the customer perceived incorrectness.

## 6 Conclusion and Future Work

In this paper, we presented CENSOR and its variants which takes <UA, SA> pair as input and generate predefined class labels targeted towards estimation of the quality of the e-commerce product data. We leveraged off-the-shelf LLM and also open-sourced pretrained LLM to showcase that fine-tuned smaller parameter model perform better in comparison. In the future, we aim to improve the performance by exploring different LLM architectures which can address the estimation of quality effectively including more modalities.

## 7 Limitations

Our work did not explore the possibility of fine-tuning Decoder-only architectures. Our experiments specifically focused on examining the encoder-decoder architecture such as Flan-T5-XL (Chung et al., 2022) and fine-tuning it using a product catalog dataset that combines human-curated data with weakly labeled data generated with an off-the-shelf LLM. This choice is primarily influenced by its recognized excellence in performance and our resource constraints. We note that our data augmentation approach by off-the-shelf LLMs and fine-tuning strategy can be used with any other pretrained Decoder-only models such as OpenLlama (Geng and Liu, 2023) and Mistral (Jiang et al., 2023).

## References

Anthropic. 2023. Model card and evaluations for claude models. *https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.*

[3] Due to organization policy, we do not report the baseline numbers and only showcase overall improvements over the baseline with proposed approaches.

Table 3: UA-specific Results.

| | UA-wise Results | | |
|---|---|---|---|
| Model | Precision (↑)<br>(Weighted) | Recall(↑)<br>(Weighted) | F1-score(↑)<br>(Weighted) |
| | Title | | |
| CENSOR-Zero-shot | +29.3 | +6.3 | +11.2 |
| CENSOR-Fine-tuned | | | |
| +Label | +36.8 | +14.8 | +29.4 |
| +Synthetic Data+Label | **+46.4** | **+21.4** | **+33.5** |
| | Bullet Points | | |
| CENSOR-Zero-shot | +19.3 | +5.1 | +14.7 |
| CENSOR-Fine-tuned | | | |
| +Label | +38.8 | +16.2 | +30.8 |
| +Synthetic Data+Label | **+44.6** | **+24.5** | **+36.3** |
| | Product Description | | |
| CENSOR-Zero-shot | +23.7 | +9.1 | +18.7 |
| CENSOR-Fine-tuned | | | |
| +Label | +33.9 | +11.5 | +26.5 |
| +Synthetic Data+Label | **+41.2** | **+19.9** | **+32.2** |

Table 4: Class-specific Results.

| | Class-wise Results | | |
|---|---|---|---|
| Model | Precision (↑) | Recall (↑) | F1-score (↑) |
| | INCONSISTENT | | |
| CENSOR-Zero-shot | -7.1 | -1.4 | -3.7 |
| CENSOR-Fine-tuned | | | |
| +Label | -0.2 | +18.7 | +9.7 |
| +Synthetic Data+Label | **+0.9** | **+31.4** | **+15.1** |
| | CONSISTENT | | |
| CENSOR-Zero-shot | +9.2 | -8.2 | +5.3 |
| CENSOR-Fine-tuned | | | |
| +Label | **+42.9** | -28.9 | +13.1 |
| +Synthetic Data+Label | +27 | **-4.9** | **+18.3** |
| | UAMS | | |
| CENSOR-Zero-shot | +6.9 | -44.9 | -19.5 |
| CENSOR-Fine-tuned | | | |
| +Label | +22.1 | **-5.3** | **+11.6** |
| +Synthetic Data+Label | **+36.6** | -20.7 | +7.8 |

Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason HD Cho, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Knowledge graph completion models are few-shot learners: An empirical study of relation labeling in e-commerce with llms. *arXiv preprint arXiv:2305.09858*.

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

Jiawen Deng, Hao Sun, Zhexin Zhang, Jiale Cheng, and Minlie Huang. 2023. Recent advances towards safe, responsible, and moral dialogue systems: A survey. *arXiv preprint arXiv:2302.09270*.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *arXiv preprint arXiv:2304.04052*.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama. *URL: https://github.com/openlm-research/open_llama*.

Jie Gui, Tuo Chen, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. 2023. A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. *arXiv preprint arXiv:2301.05712*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 17–26. The Association for Computer Linguistics.

Anitha Kannan, Inmar E Givoni, Rakesh Agrawal, and Ariel Fuxman. 2011. Matching unstructured product offers to structured product specifications. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 404–412.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Ecomgpt: Instruction-tuning large language model with chain-of-task tasks for e-commerce. *arXiv preprint arXiv:2308.06966*.

Aman Maadan, Ashish Mittal, Ganesh Ramakrishnan, and Sunita Sarawagi. 2016. Numerical relation extraction with minimal supervision. In *AAAI Conference on Artificial Intelligence*. AAAI Press.

Reza Yousefi Maragheh, Lalitesh Morishetti, Ramin Giahi, Kaushiki Nag, Jianpeng Xu, Jason Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Llm-based aspect augmentations for recommendation systems. *ICML Workshop on Deployable Generative AI*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Kartik Mehta, Ioana Oprea, and Nikhil Rasiwasia. 2021. Latex-numeric: Language-agnostic text attribute extraction for e-commerce numeric attributes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 272–279. Association for Computational Linguistics.

Ajinkya More. 2016. Attribute extraction from product titles in ecommerce. In *Workshop on Enterprise Intelligence, KDD*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

OpenAI. 2023. Gpt-4 technical report. In *arXiv:2303.08774*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pages 194–206. Springer.

Liling Tan, Maggie Yundi Li, and Stanley Kok. 2020. E-commerce product categorization via machine translation. *ACM Trans. Manag. Inf. Syst.*, 11(3):11:1–11:14.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yaqing Wang, Yifan Ethan Xu, Xian Li, Xin Luna Dong, and Jing Gao. 2020. Automatic validation of textual attribute values in e-commerce catalog by learning with limited labeled data. In *Proceedings of the 26th*

*ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2533–2541.

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.

Huimin Xu, Wenting Wang, Xin Mao, and Man Lan. 2019. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *ACL*. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wenyuan Zhao. 2020. Classification of customer reviews on e-commerce platforms based on naive bayesian algorithm and support vector machine. *J. Phys.: Conf. Ser. 1678 012081*.

Guineng Zheng, Subhabrata Mukherjee, Luna Dong Xin, and Feifei Li. 2018. Opentag: Open attribute value extraction from product profiles. In *KDD*.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2129–2139. Association for Computational Linguistics.