Towards Translating Objective Product Attributes Into Customer Language

Ram Yazdi
Amazon
ramyazdi@amazon.com

Oren Kalinsky
Amazon
orenk@amazon.com

Alex Libov
Amazon
alibov@amazon.com

Dafna Shahaf
Amazon
dshahaf@amazon.com

Abstract

When customers search online for a product they are not familiar with, their needs are often expressed through *subjective* product attributes, such as "picture quality" for a TV or "easy to clean" for a sofa. In contrast, the product catalog in online stores includes objective attributes such as "screen resolution" or "material".

In this work, we aim to find a link between the objective product catalog and the subjective needs of the customers, to help customers better understand the product space using their own words. We apply correlation-based methods to the store's product catalog and product reviews in order to find the best potential links between objective and subjective attributes; next, Large Language Models (LLMs) reduce spurious correlations by incorporating common sense and world knowledge (e.g., picture quality is indeed affected by screen resolution, and 8k is the best one). We curate a dataset for this task and show that our combined approach outperforms correlation-only and causation-only approaches.

1 Introduction

Objective catalog attributes have been part of product search for decades (Wei et al., 2013; Liberman and Lempel, 2014; Basu et al., 1998). Objective attributes are a set of pre-specified vocabulary of catalog product attributes (e.g., "price", "size", "brand", "material"), whose values ("$9.90", "wood") are provided by the sellers. These attributes are commonly used to to split search space as facets (Wei et al., 2013; Liberman and Lempel, 2014) or for product comparison (Vedula et al., 2022, 2023).

Objective attributes play an important role in defining the technical aspects of the product. However, customers tend to refer to more subjective attributes (e.g., "value-for-money", "durable", "comfortable") when referring to a product in their own language. Providing this translation to the user is becoming a key problem in product search and comparison (Radlinski et al., 2022), and is even more critical for products lacking reviews. Specifically, we outline use cases for both directions: Search engines for user searches that use subjective terms can use this translation to correctly apply the appropriate filters. Alternatively, online stores can use the translation to explain overly-technical objective attributes to users using subjective terms.

Given an objective product catalog as well as *ratings* for subjective aspects of products from the catalog, one could easily compute *correlations* between objective and subjective attributes. However, correlations might be spurious (due to chance), or they could be attributed to confounding factors. For example, a high number of HDMI ports is highly correlated with good picture quality because newer TVs tend to have many HDMI ports. Showing such spurious relations to a knowledgeable customer can lead to distrust.

One could also try to identify *causal* links between the objective and subjective attributes using recent Large Language Models (LLMs). However, while such causal links might hold true in general, they may not be applicable to the specific product catalog. The color of a shoe could potentially affect the ease of cleaning; however, for a catalog consisting exclusively of shoes made of washable materials, the impact of color on the ease of cleaning becomes marginal.

In this work, we first extract subjective and objective attribute pairs that have high correlation, based on Amazon customer rating for subjective aspects. We then apply LLM-backed causation prediction to identify promising objective-to-subjective mappings. This approach allows us to provide links that are grounded in both world knowledge as well as the product catalog. However, often, just mapping on the attribute level is not informative enough, i.e. it's obvious that the shoe material is affecting ease of cleaning, but, which material is easiest to clean?

To that end, our defined task is to both identify a causal relationship as well as the best objective attribute value. Our methods show promising results compared to several baselines on a dataset curated specifically for this novel task.

Our contributions are as follows:

- We define a novel causal-mapping task between objective catalog attributes and values and subjective, customer-driven attributes.
- We devise a solution that grounds LLMs with correlation-based methods, outperforming baselines.

2 Related Work

Subjective product attributes in recommendation systems. A common problem in natural language (Pontiki et al., 2016; Do et al., 2019; Nazir et al., 2020; Liu et al., 2020) is known as Aspect extraction is extracting these aspects and the sentiment towards these aspects, to provide a summary of these subjective attributes for each product. One example is Amazon's ByFeature Star Rating in Figure 1 that provides a rating for subjective attributes that are relevant to the product. Unlike Objective attributes that are objectively true, there is an inherent disagreement when it comes to subjective attributes. For example, a Sofa that is comfortable for one person may be less comfortable for another.

In this work, we follow the holistic definition of subjective attributes in recommender systems (RSs) by Radlinski et al. (2022). In their work, they define the three different forms of subjective attributes which we further detail in Section 3. In addition, they list different research challenges but refrain from solving the problem of search recommendation with subjective attributes. Other solutions address the problem of subjective attributes in RS in a more implicit approach. Balog et al. (2021) try to measure how soft the subjective attributes (i.e., their level of subjectivity) as to try and impact the subjective attribute rating for a given product. Zhang et al. (2014) use subjective attributes to explain why a product was recommended for a given customer based on their review. Finally, Li et al. (2019) devise a subjective attribute database to allow for search using subjective terms.

This problem can also be formulated as a vocabulary mismatch problem (Gopichand et al., 2020). Traditionally, this problem was defined as a mismatch between the user language and the document language similarly to our use case. However, existing work solve it through common approaches such as query expansion, tagging and phrase docs, yet these approaches are objective in nature and refer to the same document term in a different manner (e.g., synonyms). In contrast, we infer the relation between subjective attributes and objective attributes for a specific product type, such as "easy to clean" and the "black" color for shoes. To the best of our knowledge this problem had not been previously addressed by prior work.

Causal inference in recommendation systems. Causal inference in recommendation systems is a well-studied area (Liang et al., 2016; Wang et al., 2020; Gao et al., 2022). Existing recommendation systems learn the correlation in the data by trying to predict customer preference, better handling biased or missing data. For example, the recommender system can offer a phone charger after buying a phone, but not vice versa. Existing works utilize traditional causal inference solutions such as Structural Causal Models (Pearl, 1995) or potential outcome frameworks (Rubin, 1974).

Recent works (Kıcıman et al., 2023; Zhang et al., 2023) have evaluated modern Large Lanugage Models (LLMs) on several causal inference tasks and shown that on some tasks, these models are able to outperform traditional approaches by a large margin. The vast size of these models, together with the pretraining on the entire text on the web, allow these models to detect the causal relationship between objects to some extent.

While these works do not imply that complex causal reasoning has spontaneously emerged in LLMs, they do highlight their potential for answering causal questions that are rooted in common sense. Thus, in this work we use a recently released large language model as part of our solution to detect the causal effect between the subjective and objective attributes.

3 Problem Definition

An *objective product attribute* is a property of the product such as price, brand, size, etc. An *attribute value* is an instance of the property, such as 'blue' for the color property, or '32 inch' for the size property. We define a *subjective attribute* as any phrase or term describing the product that can be interpreted differently by two different people. For example, in the search queries "comfortable bed sheets" or "great screen quality tv", "comfortable" and "great screen quality" are subjective terms.

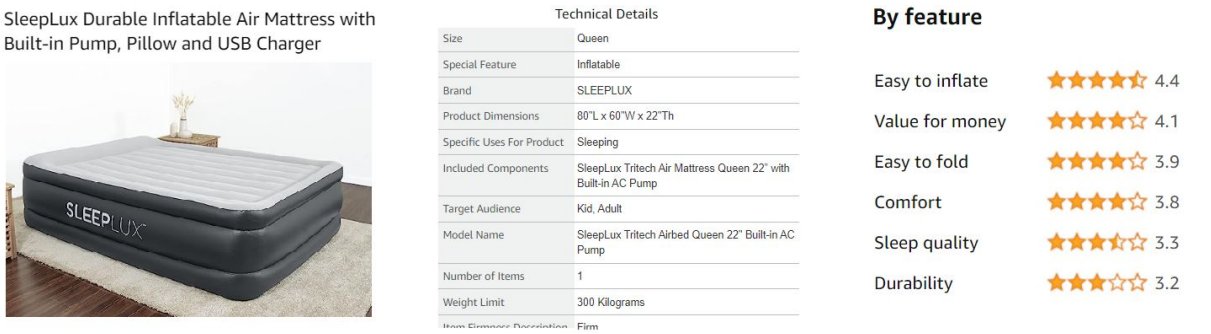

Figure 1: An example of a product (left), its objective (middle), and subjective (right) attributes in Amazon.

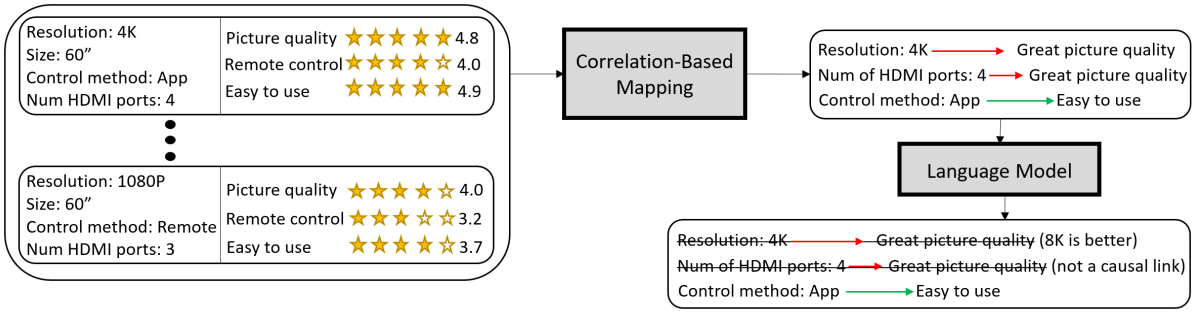

Figure 2: The pipeline takes the objective and subjective attribute values for all the products of a product type to discover correlated mappings, find the optimal objective value, and filter using a LLM to only include causal links.

However, "32 inch tv" is an objective search query since "32 inch" is a factual product attribute: the TV is either 32 inch or it is not.

In order to understand the different forms of subjectivity, we adopt the framework proposed by Radlinski et al. (2022). This framework defines three distinct forms of subjectivity:

- *Degree subjectivity* – arises when an ordinal attribute is translated into a boolean by the customer (e.g. "cheap" for price, "lightweight" for weight).
- *Compositional subjectivity* – occurs when an attribute is composed of a combination of more fundamental attributes (e.g., TV "picture quality" is mapped to "technology" and "screen resolution").
- *Semantic subjectivity* – arises when an attribute is imbued with different meanings by different customers (e.g., "funny", "cute"). Inferring personal meaning for these attributes will generally require assessment of specific items by the customer. Even product experts may disagree upon which properties lead to a "cool shirt"; one may like a cartoon design and another an impressive illustration.

In this work, we consider only two types of subjectivity – degree subjectivity and compositional subjectivity, as the third type cannot be mapped to objective facets without inherent personalization, which we defer to a later work.

We consider the following setting: Assume we are given a set of products of the same *product type* (e.g., televisions) from a catalog, each with its objective attributes and their corresponding values. In addition, a subset of the products are rated for a set of subjective attributes. In practice, such ratings can be procured through features such as Amazon's ByFeature Star Rating (Figure 1), or through analyzing review texts using Aspect-based Sentiment Analysis methods.

We define the following binary classification task: In the context of a *product type*, given a subjective attribute, an objective attribute and a value (e.g., {"Picture Quality", "Screen Resolution", "4K"}), determine whether there is a causal relation between the subjective and objective attributes, and if so, whether the value is the best option out of all the objective attribute's values. (e.g. "Screen resolution directly affects picture quality for TVs and 4K is the best resolution out of {720P, 1080P, 4K}").

4 Methodology

Our approach combines statistical correlation-based methods and Large Language Models (LLMs). We leverage the strengths of LLMs while grounding their outputs with customers feedback from the products.

Our pipeline is demonstrated in Figure 2. We are given a *product type* (e.g., a television). Then, we apply the following steps:

1. We retrieve from the catalog all specific products of the given *product type* and their corresponding *objective attributes*.
2. For each product, we also retrieve *rated* subjective attributes. We use the Amazon ByFeature attributes and their corresponding ratings (see Figure 1).
3. We apply correlation-based methods to find objective attribute values that are most positively correlated with subjective attributes (e.g., the screen resolution value that leads to the best picture quality TV).
4. We apply LLMs to eliminate objective attributes that do not directly impact the subjective attribute they are correlated with and validate the best objective attribute value selection, as detailed below.

Correlation-based methods (Step 3). We construct an indicator variable for each objective attribute value to indicate its presence or absence in a given product. We then calculate the *Point-biserial correlation coefficient* between the indicator variable and the average rating of the subjective attribute. For each subjective attribute, we select objective attribute values that exhibit a positive correlation with it.

It is worth noting that in the Amazon catalog, the sentiment towards the subjective attribute is inferred from the subjective attribute's rating, which is a continuous number at the scale of 1-5 (see Figure 1).

LLMs as an external source of knowledge (Step 4). LLMs have proven to be a powerful tool for acquiring external knowledge and capturing the complexity of natural language. When trying to establish *causal* relationships between objective and subjective attributes, LLMs can serve as an external source of knowledge to supplement traditional statistical approaches.

In our pipeline, we use the LLM to assess whether there is a direct causal relationship between the objective and subjective attribute pair, rather than mere correlation. Second, to validate using world-knowledge that the objective attribute value is the most appropriate *value* when multiple options are available (e.g., for screen resolution, 4K is indeed superior to 1080P, 720P). For each mapping identified in the previous step, we prompt the LLM, asking both whether (1) the objective attribute indeed directly affects the subjective attribute, and (2) whether value identified by our correlation-based mapping is indeed the best choice.

We use a 5-shot in-context learning setup (see Appendix B for details). These examples provide the LLM with the necessary context. We have also asked the LLM to include the reasons leading to the final answer, consistent with chain-of-thought approaches (Wei et al., 2022).

5 Experimental Study

Our goal in this section is to assess whether our method can **discover the best objective attribute value for improving a subjective attribute**.

5.1 Dataset

We chose 12 product types from diverse product categories (electronics, textile, etc.) and collected their corresponding individual products. For each product type, we identified 10 popular objective attributes (the ones used most often as filters during product search for this type). For each product type we also collected 10 subjective attributes and their ratings for individual products from Amazon ByFeature, as previously described. Altogether, this resulted in 1200 potential relationships and 54 unique subjective attributes.

For each objective-subjective relation, an annotator[1] was asked to evaluate if the mapping reflects a causal relationship independently from the catalog (see the annotation guideline in Appendix C). 17.7% of all the relations are causal links. The causal link ratio ranges between 37.5% for *Sofa* to 2.4% for *Keyboard*, showing that some products can better benefit from this mapping than others.

Next, we identify the best objective attribute value ("4K") for a certain subjective attribute ("screen quality"). We compute the average ByFeature rating for the subjective attribute, given by thousands of Amazon customers, for products with each attribute value separately. For each objec-

[1] done by domain expert employees using web search if necessary

tive attribute ("resolution"), we choose the attribute value with the maximal average ByFeature rating. We note that while this process could in theory be affected by confounders (i.e., the product groups we compare could be different in other important dimensions), we believe this is a reasonable heuristic that works well in practice.

5.2 Baseline Methods

Correlation-only. The mapping of objective attributes to subjective ones is accomplished solely through the use of Point-biserial correlation coefficients, similar to the beginning of our pipeline. Here, for each subjective attribute and objective attribute, we select the objective attribute value with the highest correlation to the subjective attribute. For the causality task, we consider any attribute that has a value with a positive correlation (i.e., correlation higher than 0) to a subjective attribute to be a causal link, which may result in many false positives.

Matching-only. Matching is a widely used approach for estimating causal effects, particularly in observational studies (Rubin, 1976). Matching methods employ a comparison between treated and control units with similar observed characteristics to estimate the effect of a treatment and address potential confounding factors. In our study, we consider each objective attribute value as a separate treatment variable and estimate the effect of that variable on the subjective attribute, while considering all other objective attributes of the product as product characteristics. Then, based on the objective attributes of each product, we match it to the most similar product, differing by the objective attribute value that is considered as a treatment. By comparing the sentiment towards subjective attributes, we are able to calculate the individual treatment effect (ITE) for each pair of products. These individual effects are then aggregated over all product pairs to estimate the average treatment effect (ATE), which serves as a measure of the causal effect. Finally, we output the objective attributes with values that have the greatest causal effect on the subjective attribute.

Formally, for each subjective attribute s and objective attribute o that is considered as a treatment we estimate the $ATE(o, s)$ as follows:

$$ATE(o, s) = \frac{1}{N} \sum_{p=1}^{N} ITE(o, s, p)$$

Method	Precision	Recall	F1
Correlation-only	8.45	7.5	7.94
Matching-only	10.6	6.25	7.87
Hybrid-Corr-Match	10.52	3.96	5.75
OpenAssistant	2.23	7.92	3.48
GPT3.5-Turbo	9.45	13.86	11.24
Corr-LLM-OA (ours)	26.78	**29.7**	28.16
Corr-LLM-GPT (ours)	**73.52**	24.752	**37.03**

Table 1: Test results of different mapping approaches.

Where N is the total number of matched pairs of products and $ITE(o, s, p)$ is the individual treatment effect of the objective attribute o on the subjective attribute s for product pair $p = (t, c)$:

$$ITE(o, s, p) = R(s, t) - R(s, c)$$

$R(s, t)$ is the average rating of the subjective attribute s for the treated product t in pair p, and $R(s, c)$ is the average rating of the subjective attribute s for the control product c in pair p. Note that treated and control products differ only by the objective attribute o.

Hybrid correlation-matching. We begin by applying the *Correlation-only* method to identify the highly correlated objective attributes per subjective attribute. We then restrict our focus to the objective attributes that were found to be correlated. Finally, we employ the *Matching-only* method to filter out any objective attributes that do not have a causal effect on the subjective attribute. By narrowing down the set of product characteristics (i.e., objective attribute), we are able to find more matching product pairs as the similarity criterion is more precise.

LLM-only. We rely solely on the LLM prediction to predict both the causality indicator and the most appropriate value of the objective attribute. Although powerful, LLMs may suffer from biases that are inherent in their training data and are not grounded on the datasets in question. Consequently, biased predictions may occur, particularly when attempting to predict the best value for the objective attribute. We consider the following language models: (a) Open-Assistant: a 12B-parameter open-source LM, (Köpf et al., 2023) and (b) GPT3.5-Turbo also known as ChatGPT.

For our own method (Section 4), we also test the performance of both OpenAssistant and GPT3.5-Turbo as the underlying LLMs, which we refer as Corr-LLM-OA and Corr-GPT-LLM, respectively.

Method	Precision	Recall	F1
Correlation-only	33.8	16.32	22.01
Matching-only	23.4	13.75	17.32
Hybrid-Corr-Match	21.05	10.0	13.55
OpenAssistant	22.01	**53.74**	31.23
GPT3.5-Turbo	**44.59**	44.89	**44.74**

Table 2: Ablation for causality only, catalog agnostic.

5.3 Results

The results are shown in Table 1. One can see that solutions that do not combine world knowledge with the grounding from the catalog lead to poor precision and recall. These solutions are unable to effectively find the best objective attribute values for a given subjective attribute. While the GPT3.5-Turbo baseline outperforms the classical solutions, the grounding to the catalog drastically increases its precision from 9.45% to 73.52% and F1 from 11.24% to 37.03% with our Corr-LLM-GPT. We also see a similar significant increase for OpenAssistant in both precision and recall when grounding it to the catalog through our Corr-LLM-OA. These results may not be sufficient to be directly shown to customers, but drastically reduces the cost of expert validation. Moreover, the results could further be improved with newer versions of LLMs.

As an ablation test, we also gauge the ability of each component to discover the catalog-agnostic causality links (described in Section 5.1). The results in Table 2 show that the correlation-only outperforms the classical causality solution. The latter is unable to find a significant number of product matchings, which in turn, leads to noisy results. The LLM-only components, also used as the causality component in our solutions, are able to find more causality links. Our analysis shows that grounding the LLMs to the product catalog leads to an increase in precision but a drop in recall. It reduces LLM hallucinations while also eliminating causality links that do not exist in the catalog. For example, the color of an apron may impact how easy it is to clean. Yet, if the catalog mostly consists of aprons from an easy to clean material, no matter the color, then there will be no correlation between the color and ease of cleaning of aprons in the catalog.

6 Observations

Below we describe two interesting phenomena we observe in the data.

Contextualized mapping. Our framework produces mappings that link objective attributes to different subjective attributes. One can see that the mapping indeed depends on the context (that is, the product type). For example, when the query is "storage box", the objective attribute "color" does not appear to have a direct impact on any of the subjective attributes, and therefore is not mapped to any of them. However, when the query is changed to "shoes", the attribute "color" is mapped to the subjective attribute "easy to clean", and for the query "measuring cup", it is mapped to "easy to read". This demonstrates the ability of our framework to create tailored mappings that are specific to the context, rather than providing a static mapping for all queries (see more examples in Table 3).

Customer expectations in product reviews. While most inconsistencies between the LLM and the correlation-based method seem to originate from the difference between the general opinion and the products available in the Amazon catalog, we find that there are a few inconsistencies that originate from customer bias. For instance, our findings indicate that metal chairs are often rated as more "lightweight" than plastic chairs, despite the latter being objectively lighter. Similarly, customers rate ashtrays made of Crystal as sturdier than those made of Metal, although Metal is generally considered sturdier.

We posit that these discrepancies can be explained by the fact that customers evaluate a product based on their preconceived expectations of it. Therefore, when evaluating a Crystal ashtray, customers may rate it as sturdy relative to their expectations of it being fragile, rather than in comparison to ashtrays made of Metal. Such divergent expectations can skew product ratings and make it difficult to make objective product comparisons. Consequently, in those rare cases, relying solely on customer feedback without considering individual expectations can lead to erroneous conclusions and hinder accurate product comparisons. Therefore, incorporating the world knowledge embedded in the LLM, such as the fact that Metal is heavier than Plastic, is crucial to account for such biases.

7 Conclusions

In this work, we define a novel task of mapping objective product catalog attributes to subjective product aspects. We show that combining correlation-based and causation-based methods (with state-of-the-art LLMs for causality) out-

Method	Query	Objective att. type	Objective att. value	Subjective att.
Corr-LLM-GPT	shoes	color	blue	easy to clean
	measuring cup	color	gold	easy to read
	ashtray	material	ceramic	heat resistance
	apron	material	polyester blend	wrinkle-free
Correlation-only	chair	material	metal	lightweight
	ashtray	material	crystal	sturdy

Table 3: Examples of objective-subjective attribute pairs mapped by our method and the Correlation-only method.

performs correlation-only and causation-only approaches. We also demonstrate that our mapping may depend on the product category, (e.g., color of shoes affects ease of cleaning, TV color does not).

As future work, we outline the problem of incorporating customer *expectations* in product reviews. We posit that customers sometimes rate subjective aspects based on expectation from the product itself rather than the broad product category, making direct comparison between ratings inaccurate. We believe that subjectivity is an under-studied area that could benefit many AI domains involving natural language, and hope this work would spur further research on this important and complex topic.

References

Krisztian Balog, Filip Radlinski, and Alexandros Karatzoglou. 2021. On interpretation and measurement of soft attributes for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 890–899.

Chumki Basu, Haym Hirsh, William Cohen, et al. 1998. Recommendation as classification: Using social and content-based information in recommendation. In *Aaai/iaai*, pages 714–720.

Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.

Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2022. Causal inference in recommender systems: A survey and future directions. *arXiv preprint arXiv:2208.12397*.

G Gopichand, S Kowshik, C Reddy, M Kumar, and P Vardhan. 2020. Vocabulary mismatch avoidance techniques. *International Journal of Scientific & Technology Research*, 9:2585–2594.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Yuliang Li, Aaron Xixuan Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and Wang-Chiew Tan. 2019. Subjective databases. *arXiv preprint arXiv:1902.09661*.

Dawen Liang, Laurent Charlin, and David M Blei. 2016. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI.*

Sonya Liberman and Ronny Lempel. 2014. Approximately optimal facet value selection. *Science of Computer Programming*, 94:18–31.

Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah. 2020. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Filip Radlinski, Craig Boutilier, Deepak Ramachandran, and Ivan Vendrov. 2022. Subjective attributes in conversational recommendation systems: challenges and opportunities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12287–12293.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Donald B Rubin. 1976. Multivariate matching methods that are equal percent bias reducing, i: Some examples. *Biometrics*, pages 109–120.

Nikhita Vedula, Marcus Collins, Eugene Agichtein, and Oleg Rokhlenko. 2022. What matters for shoppers: Investigating key attributes for online product comparison. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, pages 231–239. Springer.

Nikhita Vedula, Marcus Collins, Eugene Agichtein, and Oleg Rokhlenko. 2023. Generating explainable product comparisons for online shopping. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 949–957.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2020. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 426–431.

Bifan Wei, Jun Liu, Qinghua Zheng, Wei Zhang, Xiaoyu Fu, and Boqin Feng. 2013. A survey of faceted search. *Journal of Web engineering*, pages 041–064.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92.

A Limitations

We present a novel approach for mapping objective attributes to subjective attributes using a hybrid algorithm. However, in certain scenarios where either the LLM or the correlation-based method do not concur on the generated link, we opt to eliminate it. For instance, as outlined in section 6, customer expectations can impact their ratings of subjective attributes, potentially introducing bias to the links inferred by the correlation-based method. Unfortunately, eliminating these links leads to the loss of several valid links generated by the LLM, which cannot be grounded although being truthful.

Furthermore, our approach utilizes a constrained set of subjective attributes, limited to those that have been rated by customers, rather than encompassing all subjective attributes expressed naturally in customer reviews. Consequently, this constraint might lower the richness and diversity of the translation produced by our method. Also, in this work we focused on finding the best objective attribute value for a given subjective attribute. While this can be useful in many scenarios where the customer is interested in the best option, there may be cases where several objective attribute values may apply.

In future work, we intend to extract subjective attributes directly from open-ended reviews written by customers, which will allow for a more diverse translation. In addition, we wish to extend our framework to multiple objective attribute values, as well as support semantic subjective attributes.

B Prompt Template

For each product type, determine if the attribute value is the best option among the given attribute options with respect to the specified subjective aspect. Also, if the attribute does not directly affect the aspect, answer 'no'. Always end your answer with 'yes' or 'no'.

Product: TV
Attribute type: Resolution
Attribute value: 4k
Attribute options: 4k, 8k, 1080p, 720p
Subjective aspect: Picture quality
Answer: Among the given attribute options, 8k TVs have better picture quality than 4k TVs. The answer is 'no'.

Product: Chair
Attribute type: Material
Attribute value: Plastic
Attribute options: Iron, Plastic, Stone
Subjective aspect: Light weight
Answer: Chairs made of plastic are lighter than those made of stone or iron. The answer is 'yes'.

Product: Shoes
Attribute type: Color
Attribute value: Blue
Attribute options: Yellow, Blue, Black
Subjective aspect: Comfort
Answer: The color of the shoes does not affect their comfort, so this attribute is irrelevant to the

aspect of comfort. The answer is 'no'.

Product: Headphones
Attribute type: Water Resistance Level
Attribute value: Water-proof
Attribute options: Water-proof, Non-water-proof
Subjective aspect: Easy to install
Answer: The water resistance level of the headphones does not affect how easy to install they are, so this attribute is irrelevant to this aspect. The answer is 'no'.

Product: Headphones
Attribute type: Water Resistance Level
Attribute value: Water-proof
Attribute options: Water-proof, Non-water-proof
Subjective aspect: For working out
Answer: Water-proof headphones are more suitable for working out because they are more resistant to sweat and water damage. The answer is 'yes'.

Product: {product_type}
Attribute type: {objective_attribute_type}
Attribute value: {objective_attribute_value}
Attribute options: {objective_attribute_options}
Subjective aspect: {subjective_attribute}
Answer:

C Annotation Guideline

Your role is to evaluate if there is a direct causal relationship between a specific objective attribute and a given subjective aspect for different product queries. Specifically, given a query (e.g. TV), please determine whether the objective attribute (e.g. resolution) directly affects the subjective aspect (e.g. picture quality).

Here are some examples to guide your evaluations:

- Bedding set material and softness: A causal relationship exists, as materials like Cotton are generally softer than Polyester, for example.
- Shoe color and ease of cleaning: There is a causal relationship, as light-colored shoes may be more difficult to clean, for example.
- Chair color and sturdiness: There is no causal relationship, as the color has no impact on the sturdiness of the chair.