Can’t touch this: The effect of haptic information in automatically generated product descriptions on purchase intention.

Louk de Loijer

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Data Science and Entrepreneurship at the Jheronimus Academy of Data Science

Tilburg, Noord-Brabant
November 2021

Thesis committee:

P. Leung
K. Podoynitsyna

Abstract

Previous research has found evidence that haptic information can compensate for the inability to physically process product-related information in e-commerce environments and enhance consumers' purchase intentions. On this premise, this study examines whether zero-shot image and language models are capable of the automatic extraction of features in product images and the automatic generation of product descriptions with haptic information. To this end, the capabilities of various instruction methods of the zero-shot language model GPT-3 are used to generate product descriptions. By performing automatic and human evaluation, the few-shot and finetune instruction methods proved to generate the most readable and correct product descriptions that contain noticeable haptic information. To examine whether automatically generated product descriptions are indistinguishable from human-written product descriptions, and that haptic information in automatically generated product description enhance consumers' purchase intentions, an online experiment was performed. Despite not finding a relationship between haptic information in product descriptions and purchase intentions, the experiment demonstrated that participants fail to distinguish automatically generated product descriptions from human-written product descriptions.

Acknowledgement

This master thesis marks two years of studying the Master of Science in Data Science and Entrepreneurship at the Jheronimus Academy of Data Science. I would like to show my sincere appreciation to all the people involved in guiding me in this process. First, I would like to thank my supervisors, P. Leung, and K. Podoynitsyna, who have continuously provided valuable feedback throughout process of writing this master thesis. Second, I would like to thank J. Wind for guiding me in writing this master thesis on behalf of Squadra Machine Learning Company. Finally, I am grateful from my family and friends, who have been supporting me throughout this period.

Louk de Loijer,

Tilburg, December 1, 2021

Inhoudsopgave

1. INTRODUCTION 8

2. LITERATURE REVIEW 12

2.1 Need For Touch 12
2.2 Indistinguishability of AI-written Text 15
2.3 Zero-shot Language and Visual Models 17
2.3.1 CLIP 18
2.3.2 GPT-3 20
2.4 Human Natural Language Evaluation 23

3. HYPOTHESIS 26

4 METHODS 30

4.1 CLIP 30
4.2 GPT-3 31
4.3 Haptic Information Method 33
4.4 Product Description Evaluation Method 34
4.4.1 Automatic Evaluation Method 34
4.4.2 Manual Error Analysis 35
4.4.3 Human Evaluation Method 37
4.5 Consumer Purchase Intention Experiment Method 38

6. RESULTS 43

5.1 CLIP Results 43
5.2 Product Description Evaluation Results 47
5.2.1 Automatic Evaluation Results 47
5.2.2 Manual Error Analysis Results 49
5.2.3 Human Evaluation Results 52
5.3 Purchase intention experiment results 57

7. DISCUSSION 61

8. LIMITATIONS AND FUTURE WORK 64

9. REFERENCES 67

APPENDIX 72

Appendix A 72
Appendix B 73

List of Figures

Figure 1 18
Figure 3 22

List of Tables

Table 1 24
Table 2 24
Table 3 37
Table 4 37
Table 5 39
Table 6 40
Table 7 41
Table 8 43
Table 9 45
Table 10 47
Table 11 49
Table 12 52
Table 13 58
Table 14 60

Abbreviation List

This table describes various abbreviations and acronyms used throughout this thesis and the page on which they are first mentioned or defined.

Abbreviation	Meaning	Page
BLEU	Bilingual evaluation understudy	23
CLIP	Contrastive Language-Image Pre-Training	18
GPT-3	Generative Pre-trained Transformer	20
NFT	Need For Touch	12
NLG	Natural Language Generation	10
NLP	Natural Language Programming	9

1. Introduction

Since the invention of the Internet, e-commerce sales channels have become increasingly important for retailers to generate sales and reach new customers, resulting in a shift from physical sales to online sales (Deloitte, 2012). With forced closure of physical stores, the COVID-19 pandemic has accelerated this shift even more, and mobile and online shopping have become preferred shopping methods for most consumers (Irrera, 2021). Consequently, offering a user-friendly and convenient online shopping experience is one of the main focuses of online retailers. An important aspect of providing a user-friendly e-commerce experience is providing accurate and complete information, enabling consumers to make an informed decision. What should be provided in particular is information concerning the physical attributes of the products offered, because consumers continue to encounter obstacles as they cannot inspect the products directly to evaluate its features (Peck & Shu, 2009). In physical environments, touching the product tends to enhance consumers' ability to process product-related information and increase confidence in the purchase decision (Park, 2006). Unfortunately, physically inspecting the product is not possible on e-commerce platforms. To compensate for this, e-commerce platforms could provide *haptic information*, defined as descriptions of what touching the product would feel like. These descriptions convey a sensorial, experiential sensation, and have been proven to simulate a physical environment, which ultimately leads to increased purchasing intentions (Silva et al., 2021). More precisely, conveying a sensorial, experiential sensation causes consumers to imagine what touching the product would feel like. This imagination, referred to as *haptic imagery*, increases the perceived ownership of consumers and perceived product quality, which, in turn, increase purchasing intentions (Park, 2006; Silva et al., 2021).

Overall, the need for consumers to process haptic information before they can make an informed purchase decision varies across people and across product groups (Peck & Childers, 2003). Standardized products such as books, food, and music are considered low need for touch (NFT) products, while products such as clothing, furniture, and sporting goods are considered high NFT products (Grewal et al., 2004; Levin et al., 2005). This applies in particular to the fashion industry, because direct sensory contact with fabrics and garments provides valuable product information to make an informed choice. While apparel e-commerce players continued to capture a significant market share from brick-and-mortar shops in the period 2018-2019, the COVID-19 pandemic also forced traditional shops that had previously focused mainly on physical sales to focus on a competitive online proposition (McKinsey & Company, 2020; Orendorff, 2021). Consequently, offering accurate and complete product information is demanded in the increasingly competitive market of online fashion commerce.

To offer product information in an appealing and informative way, online retailers often hire copywriters. However, writing appealing product descriptions is a time-costly process. This makes the automatic generation of product descriptions a fast alternative capable of generating large quantities at once. Recent advancements in the field of machine learning, and notably natural language programming (NLP) have enabled the automatic generation of texts such as poems, news articles, and attempts at the automatic generation of product descriptions (Chen et al., 2019; Köbis & Mossink, 2021). Moreover, innovative models developed in the deep learning domain automatically extract features from clothing images (Wang et al., 2018).

Most of the current innovative models in feature extraction or language generation achieve state-of-the-art on a specific task. However, replicating the results of these state-of-the-art models is difficult, as the code is regularly missing, resulting in the so-called replication problem of deep learning. In fact, around 80% of published articles do not attach a code implementation (Papers with code, 2021) Another downside of these supervised task-specific models is that they are trained on huge datasets. They are often overly specific to the training distribution and do not function in other contexts. Hence, for the same task with slightly different images, these models often need to be trained again to adjust to the shift in data (Wortsman et al., 2021).

A consequence of this is the growing interest in and development of so-called zero-, one-, and few-shot machine-learning models. In contrast to traditional machine learning models, zero-shot models are generally only trained once on even larger datasets. They can be used on various tasks and generalize better on out-of-distribution tasks. The terms zero-, one-, and few-shot refer to the fact that these models require no, a single, or only a few examples as instruction for the task (Radford et al., 2021).

One of the most powerful zero-shot language generation models is GPT-3, a third-generation Generative Pre-trained Transformer, developed by OpenAI, a research laboratory for artificial intelligence backed by Microsoft. Upon its release, it received much attention in both technical and mainstream media, with coverage ranging from admiration of its capabilities to anxious warnings about potential misuse (Das, 2020). On several benchmarks, GPT-3 achieves competitive or superior state-of-the-art performance, some of which was previously held by task-specific finetuned models (Brown et al., 2020). Besides GPT-3, OpenAI also developed Contrastive Language-Image Pre-Training (CLIP), a visual language model capable of zero-shot image classification outperforming supervised state-of-the-art classifiers on multiple image classification benchmarks (Radford et al., 2021). Despite these results, zero-shot machine learning research remains a relatively new field, and both papers referenced in this paragraph acknowledged that areas exist in which these zero-shot models do not generalize well and achieve inferior performance.

Given the above, this study explores the generalizability of the zero-shot image model CLIP and the natural language generation (NLG) model GPT-3 in fashion e-commerce to automatically provide a pleasant e-commerce experience by providing accurate and complete information product information in an appealing form. Hence, this study aims to test CLIP's capabilities in product classification and feature extraction and explore GPT-3's capabilities on generating appealing product descriptions. Not only is the general quality of the product descriptions important, but to facilitate a complete e-commerce experience, they should also contain haptic information for the inability to process the products physically. If the capabilities of GPT-3 are successfully applied in the fashion e-commerce domain *with* haptic information, it allows fashion e-commerce retailers to automate the online product information offering process with user-friendly information relevant to the customer decision process. Moreover, since these models are highly novel, the application of CLIP and GPT-3 capabilities in unexplored areas is recommended to further enhance the technology and identify areas in which these models still require substantial improvement

(Brown et al., 2020; Radford et al., 2021). To validate the effectiveness of haptic information in GPT-3-generated product descriptions from a consumer's perspective, a human experiment is conducted. This experiment will test whether consistent with human-written product descriptions (Silva et al., 2021), GPT-3-generated product descriptions with haptic information similarly affect participants' purchasing intentions. As such, this paper aims to conduct cross-disciplinary research by integrating machine learning concepts such as NLP and computer vision with consumer theory that explains why haptic information empowers a user-friendly e-commerce experience.

In light of what was discussed above, this study aims to answer the following main research question:

> Can zero-shot image and language models be used to generate product descriptions with haptic information that increase consumers purchase intentions?

To this end, three sub-research questions have been formulated, which will be examined sequentially in this study.

1. Can CLIP accurately classify product categories and extract product features?

2. How can the most accurate and appealing product descriptions be generated with GPT-3?

3. Do GPT-3-generated product descriptions with haptic information increase consumer purchase intentions?

The rest of this study is structured as follows. The next chapter describes a literature review. This is followed by the hypothesis section, which formulates several hypotheses based on the related literature. After that, the methods used to test the hypotheses will be described, after which the results section outlines what has been found out in relation to the formulated research questions or hypotheses. Finally, the discussion interpretates and highlights the findings in relation to the related works.

2. Literature review

Since this study pertains to concepts from machine learning and consumption theory, several concepts need to be introduced more thoroughly. First, this section discusses key concepts from research into haptic information and NFT. Second, research into the distinguishability of AI-generated text is discussed. Third, the underlying principles of GPT-3 and CLIP are discussed in more detail. Fourth, literature on the methods used to evaluate the generated product descriptions on quality are highlighted. Since no standard method exists to evaluate text computer-generated product descriptions with user experiments, this section briefly discusses which studies have inspired the use of certain methods.

2.1 Need For Touch

The ability to touch a product is important in decision making, as it provides the consumer with information (e.g., temperature, shape, texture, and weight) about a product (Khandkar & Maurer, 2010). One of the reasons the sense of touch is effective in consumers' decision-making is that it increases the feeling of perceived ownership (Peck & Johnson, 2011; Peck & Shu, 2009). Increased perceived ownership contributes to an increase in purchase intention, a strong predictor of actual purchase behaviour (Peck & Johnson, 2011; Peck & Shu, 2009). As the ability to touch and evaluate a product is absent in online shopping, e-commerce platforms seek alternatives to bridge the gap between online and offline shopping. As regards products with sensory attributes in particular, consumers require more information to base their decision on and experience a NFT (Park et al., 2012). However, not everyone experiences the same need to touch a product before making a purchase decision. Whereas some consumers only touch products to place them in shopping carts, other consumers require more time exploring products with their hands before making a purchase decision. Peck and Childers (2003) propose that the differences in NFT stem from individual differences in accessing haptic information in the brain.

Not only do individual differences impact to what extent someone experiences NFT, but different product types also influence the NFT someone experiences. Generally, the NFT is high for products with sensory attributes such as colour, design, fabric, and fit, and as such, clothing is a high NFT product (Grewal et al., 2004; Levin et al., 2005; Peck & Childers, 2003). Although their view may be slightly outdated, as fashion-oriented e-commerce stores are indispensable today, Grewal et al. (2004) and Levin et al. (2005) state that high NFT products such as clothing are less likely to gain increased sales from online retailing compared to standardized products such as books, music, and computers. Nonetheless, retailers certainly need to put in extra effort to reduce consumer risk and persuade them.

One of the ways in which e-commerce retailers could reduce consumer risk is by providing textual haptic information. Rodrigues et al. (2017) test the effect of textual haptic information on purchase intention. Although they do not find that providing textual haptic information decreases the need to touch the product, they do find that it helps form a more realistic perception of the product. Additionally, despite failing to find a positive correlation between textual haptic information and perceived credibility of the information, they do find a positive relation between perceived credibility of the information and intention to purchase. Most importantly, their findings confirm a positive relationship between textual haptic information and online purchase intention.

Hence, the results of Rodrigues et al. (2017), mirroring results of others in the field such as Park (2006), reinforce the idea that providing textual haptic information results in increased purchasing intention. Park (2006) tests and demonstrates the working mechanism in this relationship. He finds that providing verbal information leads to haptic imagery, causing an increase in perceived product quality and eventually resulting in increased purchasing intentions. As imagining is defined as a cognitive process in which sensory information is presented in the working memory (MacInnis & Price, 1987) and imagery may involve sight, smell, taste and tactual sensations, haptic imagery is defined as the formation of mental representations of touch information (Kaski, 2002). In human-product interaction, haptic imagery refers to the process of making a representation in one’s mind of the information that would normally be attained through touching the product. Peck et al. (2013) show that haptic imagery can be used as a surrogate for touch, because *imagining* touching a product can have a similar effect on the perceived ownership as *physically* touching the product.

In addition to perceived ownership, multiple studies show that haptic imagery also exerts a positive influence on perceived product quality (Park, 2006; Rodrigues et al., 2017). Contrary to objective quality, which can be measured and verified according to pre-established patterns, perceived quality is abstract and subjective (Monroe & Krishnan, 1985). As consumers are unable to directly examine a product in e-commerce environments, visual imagery information can help stimulate the retrieval of haptic information about a product which is stored in memory. Generally, this compensatory effect leads to favourable product quality evaluations (MacInnis & Price, 1987; McKinney et al., 2002).

Building on the results of their earlier research, Silva et al. (2021) recently investigated the effect of haptic imagery and NFT in an online shopping environment. Their study reinforces Park's (2006) paradigm as the hypotheses regarding a positive relationship between haptic verbal information and haptic imagery, the positive relationship between haptic imagery and perceived product quality, and the positive relationship between perceived product quality and purchase intention are all supported. However, contrary to Park (2006), Silva et al. (2021) only find a positive relationship between verbal information and haptic imagery and not for pictorial information (operationalized with zoom-in option on product images). Their explanation for not finding this relationship is the use of a simple grey hoodie as a product in their consumer study vis-à-vis the fancier clothes such as dresses used in Park's (2006). For more simple clothing, a zoom-in option may evoke less memory cues as there are less details to show when zooming in.

To summarise, not only do individual differences in NFT exist, but different product groups also affect the NFT someone experiences. Personal differences regarding NFT can be measured on the NFT scale developed by Peck and Childers (2003). Regardless of the individual differences, providing haptic information leads to increased perceived ownership and increased perceived product quality, resulting in increased purchasing intentions.

2.2 Indistinguishability of AI-written Text

Studies such as the research of Rodrigues et al. (2017) and Park (2006) focus on consumer behaviour towards human-written haptic descriptions. Since this study also includes product descriptions stemming from machine learning models, the perception towards AI and consumers' ability to distinguish AI-generated text from human-written text must be considered.

Regarding the perception towards AI, ethical concerns have become more common over the years, resulting in an overall negative perception towards AI (Fast & Horvitz, 2017). More specifically, Jakesch et al. (2019) research AI-mediated communication in Airbnb profiles and find that trustworthiness decreased when participants suspected that a profile description is generated by AI. Graefe et al. (2018) find contradicting results in similar research in which participants rated computer-written articles. They find that participants tend to rate articles declared as human-written more favourably, regardless of the actual source. However, participants rate actual computer-written articles as more credible and higher in journalistic expertise but less readable.

Notwithstanding these findings, the negative perception towards AI-generated text only counts as a problem if those who are presented with it can distinguish it from a human-written text. With the development of models such as GPT-2 and GPT-3, the task of distinguishing AI content from human-generated content becomes increasingly difficult, if not impossible. As Brown et al. (2020) recognize the potential dangers of the indistinguishability of the texts generated by GPT-3, they decide to measure human ability to distinguish GPT-3-generated news articles. For this, they select 25 online news article titles, subtitles, and text (with a mean length of 200 words) and generate completions to the articles' titles and subtitles with GPT-3. A human-out-of-the-loop method is used to automatically select the completion whose word-count was closest to the original article, which was done to prevent human cherry-picking and minimize the effect that completion length might have on participants' judgments. Testing the capabilities of multiple GPT-3 models varying in size, they present participants with a quiz consisting of the real titles and subtitles followed by either the human-written or the model-generated article. Participants would then indicate to what extent they think that the article was written by a human or by a machine. By experimenting with model size, Brown et al. (2020) find that mean accuracy in detecting whether an article was human-written or GPT-3-generated shrank from 76% for the small model to 52% for the 175 billion parameter model, which is barely above chance. Moreover, they also find that participants spent more time trying to identify whether an article is GPT-3-generated as model size increases, signifying that the task becomes increasingly difficult as the model generates more sophisticated completions. They also run an experiment with articles of approximately 500 words, and mean human accuracy remains at 52% for the 175 billion parameter model. This indicates that, independent from output length, GPT-3 continues to output text which is difficult to distinguish from human-written text.

In a similar project, Köbis and Mossink (2021) test whether readers can distinguish GPT-2-generated poems from human-written poems and which of these two the readers prefer. From the generated texts, either the best poem (human-in-the-loop technique) or a random poem is sampled (human-out-of-the-loop technique). In their monetary incentivized experiment, a participant (judge) seeks to determine which of the two presented poems is machine-generated and which is human-written. Their analysis reveals that in the human-out-of-the-loop treatment, the participants' accuracy in detecting human-generated poems deviates significantly from chance, while they do not in the human-in-the-loop treatment. Their findings suggest that people are unable to reliably distinguish human-written from artificially generated poems if a human is involved in cherry-picking the best poems. In accordance with Jakesch et al. (2019) and Graefe et al. (2018), they also find that participants express a slight aversion to algorithm-generated poetry, irrespective of whether participants were informed about the algorithmic origin of the poem.

A similar study by Gunser et al. (2021) also examines whether participants with a literature-specific professional background can distinguish poems and short stories written with the help of a GPT-2-based interactive interface from texts written without this tool. Although the participants correctly classify AI-based continuations as such in approximately 81 percent of the cases, they misclassify 18% of the cases as either human continuation or the original text. These findings show that nowadays, not even professionals are perfectly capable of distinguishing between AI-generated and human-written texts.

2.3 Zero-shot Language and Visual Models

As discussed earlier, this study aims to analyse the capabilities of zero-shot language and visual models to automatically generate appealing product descriptions. Generally, the use of zero-shot image entails multiple advantages. First, zero-shot models generalize much better to out-of-data classes vis-à-vis task-specific models, which means that even on previously unseen classes, accurate predictions can be made (Wang et al., 2019). Second, attaining labelled data is notoriously difficult and is often not available due to labelling cost, domain shift, rare classes, or a vast number of total classes (Wang et al., 2019). As a matter of fact, due to labelled data scarcity, people have unknowingly been performing data labelling tasks for Google’s machine learning purposes for years (Maruzani, 2021). Although most zero-shot models are initially trained on petabytes of labelled data, they do not require class labels for unseen classes. Unfortunately, the wider generalizability comes at the expense of performance on fine-grained tasks, since no model can achieve the best generalizable classification results on all classification tasks. Although some zero-shot models can be retrained to fit a new task, this violates the principle of zero-shot learning. Transfer learning is the domain where knowledge contained in the source domain and source task is transferred to the target domain for learning a model to perform the target task.

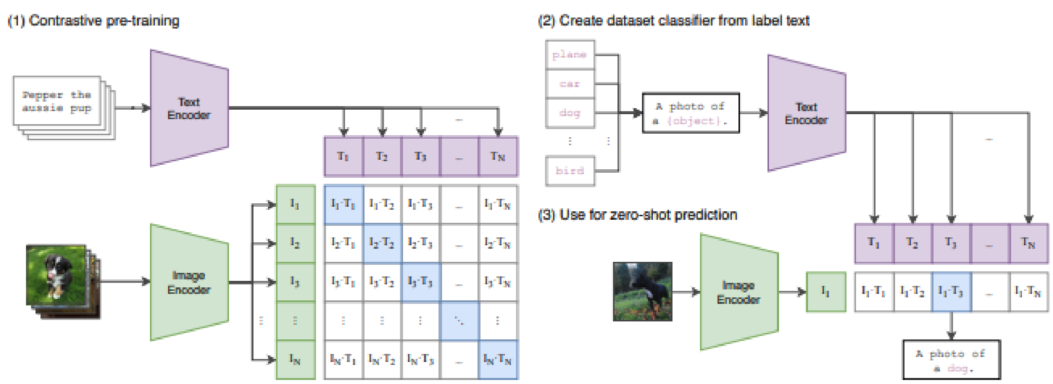
In e-commerce context, a task-specific category classification model trained on a dataset containing clothing will never be able to accurately classify shoes if not additionally trained. A zero-shot model initially trained on petabytes of data inherently ‘knows’ what a shoe looks like. Moreover, zero-shot models are less prone to domain shift, which happens frequently in fashion as products and styles change from season to season. This applies in particular to CLIP and GPT-3, the zero-shot visual and few-shot language models used, which is why the following section elaborates on these models and discusses what makes them revolutionary.

2.3.1 CLIP

Developed by the AI research and deployment company OpenAI, whose goal is to "ensure that artificial general intelligence benefits all of humanity", CLIP is a novel visual language model (Radford et al., 2021). While standard image models train an image feature extractor, and a linear classifier, CLIP trains an image encoder and a text encoder simultaneously on predicting what text corresponds to a certain image and vice versa. When CLIP is used for zero-shot prediction, descriptions of the target classes are embedded, and the description with the highest similarity to the embedded image vector is predicted as the correct label.

Figure 1

The training method of CLIP visualized



Note. from "Learning Transferable Visual Models From Natural Language Supervision", by Radford et al. (2021), p2.

While a zero-shot CLIP classifier outperforms a supervised residual neural network (ResNet)-50 classifier on 16 out of 27 datasets it was benchmarked on, it only achieves 88% accuracy on the handwritten digit MNIST data. In fact, CLIP is outperformed by a simple baseline of logistic regression on raw pixel. In an attempt to find an explanation for this reduced performance, Radford et al. (2021) verify that there are almost no images that resemble MNIST digits in the training dataset. This suggests that CLIP does little to address the problem of task-specific models not generalizing to other tasks. Instead, a reliance on

CLIP amounts to a circumvention of the problem and the assumption that, by training on such a large and varied dataset, all data will be effectively in-distribution. This is a naive assumption that, as MNIST demonstrates, is easy to debunk. As CLIP works through an embedding of the names or descriptions of the target classes, modifying the textual embeddings can increase the performance of CLIP significantly. This so-called prompt engineering, where for example the label "A photo of a [dog type label]" is changed into "A photo of a [dog type label], a type of pet", resulted in a 5% increase on ImageNet accuracy.

For the image encoder component of CLIP, Radford et al. (2021) consider the traditional ResNet architecture and the Visual Transformer (ViT), experimenting with varying sizes of both architectures. Starting from an improved version of the ResNet-50 architecture (He et al., 2016b), they upscale the depth, the width as well as the input resolution with EfficientNet scaling technique, resulting in models with roughly 4x, 16x, and 64x computation. Unlike the ResNet architecture, the ViT addresses the computer vision task with a transformer-based approach, which gained increasing recognition through its use in NLP tasks (Dosovitskiy et al., 2020). The ViT model works by flattening patches from splitting an image, producing lower-dimensional linear embeddings from the flattened patches, adding positional embeddings to these patches, and feeding the sequence as an input to a standard transformer encoder. Dosovitskiy et al. (2020) show that ViT achieves competitive performance compared to state-of-the-art convolutional networks, while requiring substantially fewer computational resources for training. The ViT bases its configurations on those used for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), which means that the model variations differ in terms of model size and input patch size. For example, the ViT-B/32 refers to the Base model with 32x32 input patch size. As the ViT's sequence length is inversely proportional to the square of the patch size, models with smaller patch size are computationally more expensive. Overall, Radford et al. (2021) test four variations of the CLIP-ViT architecture.

Combining both architectural configurations leads to nine configurations: ResNet-50, ResNet-101, ResNetN50x4, ResNet50x16, ResNet50x64, ViT-B/32, ViT-B/16, ViT-L/14, and ViT-14-336px. Across 27 benchmarks, the largest and most advanced models, ViT-L/14 and ResNet50x64, achieve the best performance on most of the datasets. The ViT-L/14 achieves the best performance on 18 datasets, while the ResNet50x64 achieves the best performance on 5. Unfortunately, these two architectures have not been published publicly, as only the smaller configurations are available when importing the CLIP library in python.

2.3.2 GPT-3

GPT-3 is a third-generation, autoregressive language model based on the transformer architecture. In July 2020, beta access was made available via an API, as OpenAI refrained from public release due to concerns about malicious applications.

Brown et al. (2020) train 8 different model sizes, ranging from 125 million to 175 billion parameters. They refer to the 175 billion parameter model as GPT-3, as this model achieved the best performance. At publication, GPT-3 contained a tenfold of the parameters of any previous non-sparse language model. The model was trained on 300 billion tokens (approximately 225 billion words), with Wikipedia, Books, WebText2, and Common Crawl data used to create a dataset large enough to train this model on. A natural consequence of the size of the dataset and the number of parameters of the model is the energy usage required to train the model. Although Brown et al. (2020) hint at the energy used to train GPT-3, no actual amount is given, but estimates state that it costs around 500MwH and $12 million in compute credit to train. This means that the model makes a significant environmental impact, which should be considered in today’s age of global warming. At the same time, the model only requires to be trained once, while supervised task-specific models need to be trained repeatedly for new out-of-scope tasks.

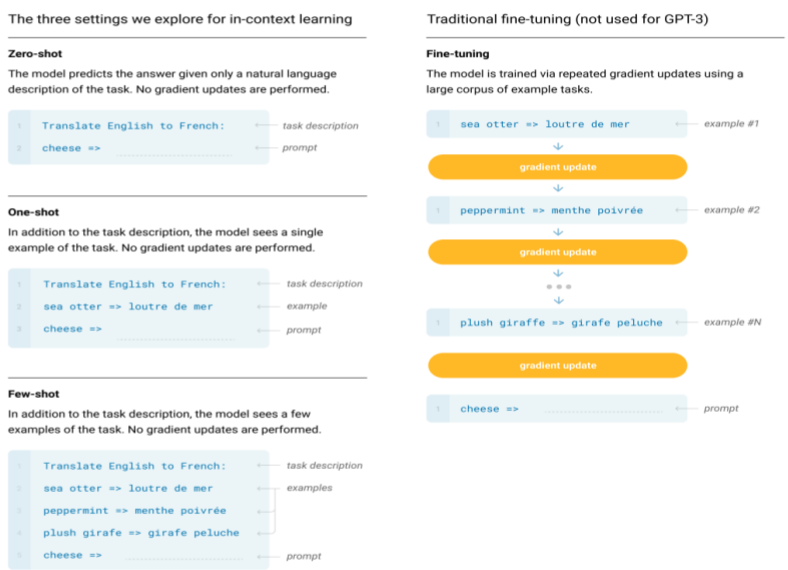
The GPT-3 model can be accessed via the GPT-3 API. The GPT-3 API works on a text-in-text-out basis, in which the input text is defined as the ‘prompt’ and the model will output text as a ‘completion’ to the prompt. GPT-3 processes text by breaking it down into tokens, which are words or chunks of characters. For example, the word ‘ecommerce’ is divided into the parts ‘ec’, ‘com’, and ‘merce’, but a word like ‘text’ remains a single token.

Since GPT-3 is a text-in-text-out model, there are multiple methods in which task-specific data is used to learn a task. Specifically, four methods are defined, depending on how many examples are provided: zero-shot, one-shot, few-shot, and finetune learning. First, there is zero-shot learning, which is essentially not providing any example and simply giving the model an instructing like ‘translate English into French: cheese =’. Given that no examples are provided to the model, this is the most challenging method that invariably achieves the lowest performance vis-à-vis the other methods. The second method is one-shot learning, in which only one example is given besides the natural language description of the task. The reason that Brown et al. (2020) distinguish one-shot from zero-shot learning is their belief that the former most closely resembles the way in which humans are instructed to perform a task. The third method, a few-shot learning approach, provides GPT-3 with several examples of the task at inference time as conditioning, but no weights are updated on these examples.

The final method in which GPT-3 can learn a desired task is finetuning. In traditional machine learning, fine-tuning involves updating the weights of a pre-trained model by training on a supervised dataset with typically thousands to hundreds of thousands labelled examples of the desired task. Frequently mentioned disadvantages of this method are the poor generalization to out-of-distribution tasks and the necessity of a large new dataset for each new task. Initially, Brown et al. (2020) do not report fine-tuning results as they focus on task-agnostic performance. Consequently, the fine-tuning method was not included in the initial beta access of GPT-3. However, in an update in 2021, OpenAI embedded this method into the API and made it available to GPT-3 beta access users. GPT-3 fine-tuning works by providing a JSON file with examples of input prompts and desired completions. Next, a finetuned version of GPT-3 is training on the provided examples. This means that finetuning is the only instruction method that involves gradient updates. Once a model has been finetuned, no examples need to be provided in the prompt anymore. Unlike traditional machine learning finetune methods, GPT-3 finetune does not require thousands to hundred thousand of examples. Instead, OpenAI suggests providing at least a hundred high-quality examples, ideally vetted by human experts. From there, performance tends to linear increase with every doubling of the number of examples.

Figure 2

Instruction methods of GPT-3



Note. from "Language Models are Few-Shot Learners", by Brown et al. (2020), p7.

Overall, on NLP tasks, GPT-3 achieves promising results with the zero-shot and one-shot methods. In fact, with the few-shot method, GPT-3 is mostly competitive with or sometimes even exceeds state-of-the-art performance, with some state-of-the-art performance being held by task-specific finetuned models. For example, on the Penn Treebank evaluation dataset for language modelling, GPT-3 sets a new state-of-the-art by a margin of 15 points, achieving a perplexity of 20.5. On LAMBADA, a dataset on the modelling of long-range dependencies, Brown et al. (2020) achieve an 86.4% accuracy with a few-shot approach, advancing the state-of-the-art by 18 percentage points. For the product description generation task, this significant increase in modelling long-range dependencies implies that GPT-3 is more capable of understanding that 'its' in the sentence 'its materials are cotton and polyester' refers to the product category, regardless of the number of sentences between the product category sentence and the material sentence.

2.4 Human Natural Language Evaluation

For automatic measures, standard reference-less metrics such as BLEU and ROUGE are the most prevalent (Lin, 2004; Papineni et al., 2002). However, automatic measures often do not correlate with human evaluations and are sometimes uninterpretable (Van der Lee et al., 2019). Due to these limitations, human evaluation remains a good standard for the assessment of overall text quality, but is commonly forgotten (Van der Lee et al., 2019). Human evaluation of NLG models can be done in many ways, which is why no standard evaluation model exists. Since no standard human evaluation procedure exists, this section discusses literature on which this study's human evaluation is based.

In a literature review, Van der Lee et al. (2019) show that 42 different criteria are used in 88 papers on human evaluation practices published in INLG and ACL. Fluency, naturalness, and quality are the three most-used criteria, appearing in 13, 8, and 5 papers, respectively. In a supplementary attachment, Van der Lee et al. (2019) provide a detailed overview of the investigated papers. Their sample contains fifteen papers which conduct human evaluation of data-to-text generation models. Out of these fifteen papers, three only conduct automatic evaluation (mostly BLEU), seven use both automatic and human evaluation, three exclusively use human evaluation, and two use neither. Concerning the participants recruited to evaluate NLG models, a distinction can be made between expert-focused and reader-focused evaluation. In the former, a small number of expert annotators are recruited to review outputs of the NLG model. The latter design typically involves a larger sample of (non-expert) evaluators.

In terms of measuring the criteria, Van der Lee et al. (2019) advise the use of multiple-item (7-point) Likert scales as findings show that 7-point scales maximize reliability, validity, and discriminative power (George A Miller, 1956; Preston & Colman, 2000). Following the advice of Van der Lee et al. (2019) and motivated by the lack of resources which makes it impossible to conduct an expert-focused evaluation, inspiration is drawn from papers that conduct reader-focused human evaluations with the use of Likert scales. Besides the papers included in the supplementary attachment of Van der Lee et al. (2019), one additional paper which conducts reader-focused human evaluation on a data-to-text-generation model is discussed.

In the research of Van der Lee et al. (2018), neural machine translation is explored for conversion from data to text. Besides BLEU-based evaluation, each of the 24 participants were asked to rate four sentences each. For participants to provide ratings on correctness, participants are provided with a table containing the information used to generate the sentences. The participants judge the quality based on the criteria shown in table 1.

Table 1

Human evaluation criteria (Van der Lee et al., 2018)

Criteria	Question
Fluency	This text is written in proper Dutch.
	This text is easily readable.
Clarity	While reading, I immediately understood the text.
Correctness	This report does not include extraneous or incorrect information.
	This report does not omit important information.

Note. Adapted from "Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods", by van der Lee et al. (2018), Proceedings of the 11th International Conference on Natural Language Generation, p40.

In the research of Qader et al. (2018), automatic generation of persuasive messages in the context of short company descriptions is tested. In their human-based evaluation, a web-based experiment is conducted with a group of nineteen users who judge the quality of the texts based on five statements on a 5-point Likert scale shown in table 2. Qader et al. (2018) state that their questions were specifically designed to measure the weakness of NLG models: content choice, repetition, hallucination, and poor segmental connection.

Table 2

Human evaluation criteria (Qader et al., 2018)

Criteria	Question
Information Coverage	How do you judge the Information Coverage of the company summary?
Non-redundancy	How do you judge the Non-Redundancy of Information in the company summary?
Semantic Adequacy	How do you judge the Semantic Adequacy of the company summary?
Grammatical Correctness	How do you judge the Grammatical Correctness of the company summary?

Note. Adapted from "Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation.", by Qader et al. (2018), Proceedings of the 11th International Conference on Natural Language Generation, p261.

As the publication of van der Lee et al. (2019) preceded the publication of Chen et al. (2019), their paper was not included in their literature review. Nonetheless, the paper is relevant to this study as Chen et al. (2019) conduct a human evaluation experiment in the context of product description generation in e-commerce. In their experiment, participants score the generated product descriptions on fluency, diversity, and overall quality on a 5-point Likert scale. However, Chen et al. (2019) do not report a sample size and neither motivate the use nor the specific questions stated for the criteria. They merely state that “The score of fluency reflects how fluent the description is”, “The score of diversity reflects how diversified the description is and whether it contains much competitive content”, and “The score of overall quality reflects whether the description is reasonable, or to say, whether it is consistent with world knowledge” (Chen et al., 2019, p. 3047).

Going back to the best practices for the human evaluation of automatically generated text, Van der Lee et al. (2019) state several important other issues which are currently overlooked in existing NLG human evaluation studies. One of these is the common use of single-item scales for one criterion instead of multiple-item scales. The main arguments against the use of a single-item scale stated are that they are unlikely able to capture a complex concept (such as fluency), do not provide enough points of discrimination, and no reliability measure for internal consistency (e.g. Cronbach’s alpha) can be calculated for a single item. Moreover, they warn that within-subjects designs are susceptible to order effects as annotators can change their responses due to fatigue, practice, or carryover effects.

3. Hypothesis

Based on the literature review, eight hypotheses are formulated addressing different aspects related to the three research questions. This section discusses how each hypothesis was established, as well as the formal formulation of the hypothesis.

To test the product classification and feature extraction capabilities of CLIP in a fashion e-commerce context, CLIP is benchmarked against a supervised finetuned model. Based on the results described in the study of Radford et al. (2021), the expectation is that CLIP performs significantly better than a supervised task-specific model. Moreover, if CLIP's performance transfers to the fashion category and attributes classification task, there likely have been a broad number of similar examples in the training data. As none of the 27 datasets on which CLIP's performance was benchmarked focus specifically on fashion-related datasets, testing the robustness on fashion classification and feature extraction falls in the recommended future works described by Radford et al. (2021). In their recommendations, they promote community exploration to further characterize the capabilities of models like CLIP to identify both application areas where these models have promising performance and areas where they have reduced performance. Based on the findings of Radford et al. (2021), hypothesis 1 states:

> **H1:** CLIP outperforms a supervised finetuned model on apparel category prediction and feature extraction.

To evaluate the quality of the product descriptions of the GPT-3 instruction methods, the product descriptions are assessed on readability and lexical diversity metrics. The four different ways of instructing GPT-3 (zero-shot, one-shot, few-shot, and finetune) are evaluated with these metrics. In the evaluation of GPT-3, Brown et al. (2020) see significant improvement in performance when providing GPT-3 with more examples. Consequently, the expectation is that the quality of the GPT-3-generated product descriptions increase when the model is provided with more examples:

H2A: Finetune and few-shot GPT-3-generated product description score better on readability compared to zero-shot and one-shot GPT-3-generated product descriptions.

H2B: Finetune and few-shot-generated GPT-3-generated product description score better on lexical diversity compared to zero-shot and one-shot-generated GPT-3-generated product descriptions.

Besides automatic evaluation, a human evaluation study is established and distributed. In this study, human participants will judge the quality of the product descriptions based on readability, correctness, and repeated information. Substantiated by the findings of Brown et al. (2020), it is expected that human evaluators rate the GPT-3-generated product descriptions instructed with more examples higher than GPT-3-generated product descriptions instructed with fewer examples:

H3: Participants evaluate finetune and few-shot-generated product description higher on readability, correctness, and repeated information compared to zero-shot and one-shot-generated product descriptions.

Moreover, the human evaluation study is used as a validation and manipulation check for the consumer experiment. Since participants will be exposed to product descriptions with and without haptic information, the way in which GPT-3 generates product descriptions by including or omitting haptic information needs to be validated. By asking participants to what degree a product description contains words that describe what the product feels like, it is validated that both haptic and non-haptic product descriptions are perceived as such. This is formulated in hypothesis four:

H4: Participants observe that haptic product descriptions contain significantly more haptic information compared to non-haptic product descriptions.

As several studies have shown that haptic imagery increases perceived product quality and perceived ownership, which increase purchase intentions, this study assumes the existence of the mediating relationships and does not aim at confirming these findings. Hence, only the direct effect of haptic information on purchase intention is measured. However, as Silva et al. (2021), contrary to literature such as Park (2006), do not find a moderating relationship for NFT on purchase intention and perceived product quality, this study aims to contribute new evidence on the moderation of NFT by examining its effect. Based on the findings of Park (2006), Silva et al. (2021), and Rodrigues et al. (2017), the following hypotheses are formulated:

> **H5:** Participants experience a higher intention to purchase when viewing an apparel product with a haptic product description vis-à-vis an apparel product without a haptic product description.

> **H6:** NFT moderates the relationship between haptic information and purchase intention.

Although the findings of Gunser et al. (2021) and Köbis and Mossink (2021) show that people already have difficulty distinguishing AI-generated from human-written text, both studies only use GPT-2, the predecessor of GPT03, which was published in early 2019. This study uses GPT-3 for product description generation, which is seventeen times larger with 175 billion parameters. Compared to news articles and poems, a product description of approximately 40-50 words is quite short, which means that it should be easier to generate flawless indistinguishable texts. However, the shorter texts also mean that participants can devote extensive attention to each small detail in syntax, semantics, and grammar. This means that a single small mistake in the product description could lead to a participant suspecting that a text is algorithm-generated. Moreover, as the product descriptions are shown besides images, a minor disparity between the image and the product description could also lead to participants tending to doubt the origin of the text. Factual inaccuracies pose a high risk to cause a disparity between the product description and image, as GPT-3 is known to hallucinate irrelevant information (Brown et al., 2020; See et al., 2017; Zhao et al, 2021). Concludingly, generating indistinguishable texts in the context of the product descriptions that accurately reflect the product image remains a challenge and may require more effort than simply instructing the prompt of GPT-3. Nonetheless, based on the impressive results reported in Brown et al. (2020), Gunser et al. (2021), and Köbis and Mossink (2021), it is expected that participants cannot distinguish GPT-3-generated from human-written product descriptions:

> **H7:** Participants are unable to distinguish GPT-3-generated from human-written product descriptions.

As it is the expectation that GPT-3 can generate product descriptions that are indistinguishable for consumers, this also means that the inclusion of haptic information in these product descriptions will be similar. Hence, if the GPT-3-generated product descriptions are in fact indistinguishable from the human-written, the mechanism of haptic information that eventually leads to increased purchase intentions will work in a similar matter. Consequently, participants are expected to respond to the GPT-3-generated product descriptions similarly as they would respond to a human-written product description. Therefore, the following and final hypothesis has been formulated:

> **H8:** Participants express at least the same purchase intention when viewing a GPT3-generated apparel product description vis-à-vis a human-written product description.

4 Methods

Since this study covers eight hypotheses addressing different aspects related to three research questions, varying methods are used to support or refute the hypotheses. This section describes the various methods used in this study. First, the method used to evaluate the capabilities of CLIP in category prediction and feature extraction are discussed. The second subsection discusses the method used to generate product descriptions based on various instruction methods. Since the generated product descriptions should contain haptic information, the third subsection discusses the method used to manipulate GPT-3 to generate haptic product descriptions. The fourth subsection covers the methods of the human evaluation, automatic evaluation, and manual error analysis performed on the generated product descriptions. The fifth subsection discusses how the consumer experiment was set up that measures the effect of haptic information on purchase intentions.

4.1 CLIP

To test the capabilities of CLIP in the domain of apparel classification and apparel attribute detection, the DeepFashion dataset is used (Liu et al., 2016). DeepFashion is a large-scale apparel dataset with a collection of 290 thousand fashion images containing 52 different classes and annotated with 1,000 attributes. Since its publication in 2016, the DeepFashion dataset has become the benchmark for evaluating various fashion-related tasks, for example category and attribute prediction. For category prediction, the top-k classification accuracy is commonly used as an evaluation metric, with k equals three and five predominantly mentioned in papers benchmarking on DeepFashion. For attribute prediction, the top-k recall rate is proposed as the evaluation criteria, which is obtained by ranking the 1,000 attribute prediction scores and calculating the number of matched attributes in the top-k predictions. Like the classification task, k equals three and five are generally used in papers benchmarking on DeepFashion. Although DeepFashion is commonly used to benchmark the performance models, classes are unbalanced, with the classes dress, tee, blouse, and shorts representing more than fifty percent of all classes. In order to prevent the imbalance from affecting the results, a more balanced subsample of the

DeepFashion dataset was created, consisting of 500 images. The balancing was not performed for training purposes, because zero-shot models do not require training data, but to ensure that the models perform as well on minority classes as on majority classes.

To determine which out of the six available model variations of CLIP best suits the apparel product classification and attribute extraction tasks, all available models are tested on the DeepFashion category classification task. Since Radford et al. (2021) find significant improvement with prompt engineering, the effect of prompt engineering is also tested in the product classification and attribute extraction task. To prompt engineer the labels of the DeepFashion attribute data, the product category is added to the attribute label. For example, the attributes ‘denim’, ‘distressed’, and ‘faded’ are embedded as ‘denim jeans’, ‘distressed jeans’, and ‘faded jeans’. Following Radford et al. (2021), For category prediction, the embedding as proposed by Radford et al. (2021) is used, resulting in for example the embedding ‘jeans, a type of clothing’.

4.2 GPT-3

GPT-3 can be instructed in various ways. “Language Models are Few-Shot Learners”, the title of Brown et al. (2020), implies that the method used to instruct GPT-3 influences its performance on various language tasks. To assess the quality of GPT-3 in product description generation, all possible instruction methods are compared, starting with zero-shot instruction up to fine-tuning. Ideally, a single dataset would be used for testing the capabilities of both CLIP and GPT-3. However, the DeepFashion dataset does not contain product descriptions required as examples in the instruction methods. Moreover, no data-to-text fashion product description dataset is publicly available. Consequently, the apparel website Nordstrom is scraped and for various product categories, product data and product descriptions are obtained. The product data and descriptions are used as examples in the prompts of the one-shot and few-shot instruction, while the finetune model is trained on a dataset of 100 pairs of product data and product descriptions.

For each instruction method, a standard prompt was written, to which the attributes of the to -be-generated product description could be added. OpenAI recommends providing the data to the GPT-3 prompt in a natural language form instead of a data format. Hence, a simple transformation (e.g. ‘colour=Blue’ to ‘the colour is Blue’) transforms all product attributes into natural language. For the zero-shot instruction, only the task description (instruction) is provided along with the attributes of the to-be-generated product description. For the one-shot instruction method, the task description is provided along with a single example. For the few-shot instruction method, GPT-3 is given a similar prompt to the one-shot instruction, but instead, three examples are provided. Hypothetically, many more examples could be provided for this instruction method (up to 2,049 tokens and approximately 1,800 words). However, as the cost of a GPT-3 API-call is proportional to the number of words in the prompt, providing three examples was regarded as the optimal setting with regard to the price versus output quality trade-off.

The finetuning process differs significantly from the other instruction methods and consists of providing GPT-3 with examples (preferably at least a few hundred vetted by human experts) of input prompts and desired outputs. Although this method more closely resembles traditional task-specific machine learning techniques, as the finetuned model is aimed to perform well on the specific task it is finetuned on, GPT-3 finetune does not require thousands to a hundred thousand examples. Currently, the finetuning process is free of charge (in the beta-access), and no examples need to be provided in the prompt once a model has been finetuned. This means that the (promised) higher quality output is offered at a cheaper price because finetune prompts, in which only the to-be-generated product data must be provided, are significantly cheaper than a token-heavy few-shot instruction. After experimentation with the finetune method, the best performing finetuned model is a model trained on a dataset of 100 examples scraped from Nordstrom.

Table 1 in appendix B shows the examples provided to the prompt for the one-shot and few-shot methods, while a sample of the input prompts and desired completions is shown for the finetune method. The zero-shot, one-shot, and few-shot completions are all generated with GPT-3’s most powerful model, Davinci, while the Curie model is used for finetuning.

4.3 Haptic Information Method

Since the aim of this study is to test the effect of haptic information in GPT-3-generated product descriptions on purchase intention, GPT-3 needs to be able to generate both haptic and non-haptic product descriptions. For this, a standardized list of 306 touch-related adjectives, developed by Stadtlander and Murdoch (2000), is used. In terms of generating a product description with a GPT-3 API request, the logit bias parameter can be provided. The logit bias parameter changes the weights of certain tokens and modifies the likelihood of the provided tokens appearing in the completion. The logit bias parameter accepts the weight of a token between -100 and +100. Generally, values between -1 and 1 decrease or increase the likelihood of selection, whereas values like -100 or 100 result in a ban or exclusive selection of the relevant token. Hence, for product descriptions without haptic information, the standardized list was provided with weights of -100 for all words. For product descriptions with haptic information, it was found through experimentation that a weight of +2 for all words resulted in the right number of touch-related words in the product description. A considerable limitation in this method is the fact that the logit bias parameter only supports words that encode to a single token. This means that touch-related words that encode to multiple tokens cannot be automatically omitted from the generated product descriptions. Some of the words that are embedded as multiple tokens are 'supersoft', 'comfort', 'stretchy', 'lightweight', 'fit', 'material', and 'ultra-tight'.

After experimentation with the logit bias parameters and inspection of the outputs, it became apparent that some product descriptions contained these multiple-token touch-related words. Especially the finetuned model, which is trained on examples with haptic information, often circumvents the logit bias method and samples multiple-token touch-related words in the product descriptions. As a result, the intended non-haptic product descriptions still contain touch-related words that may trigger haptic imagery for consumers who read these product descriptions. Since these words are contextually embedded in a sentence with the words around them, automatically replacing the touch-related words with more generic words is not possible. With the implementation of exposing participants to haptic and non-haptic product descriptions as a focus, the only solution is to manually rewrite the haptic parts of the sentence so that a non-haptic product description no longer contains touch-related words. This means the method used to manipulate GPT-3 into automatically generating haptic and non-haptic product descriptions is unsuccessful, and the product descriptions without haptic information are no longer generated fully automatically.

4.4 Product Description Evaluation Method

4.4.1 Automatic Evaluation Method

Given that methods such as BLEU and ROUGE are based on word-overlap-based metrics, they require a reference base for computing the score. Since no or a too small reference corpus exists for zero-, one-, or few-shot instructed NLG models, the output of these models must be evaluated with reference-less evaluation methods. One of the most common used evaluation methods of the quality of written text is readability, which is defined as the ease with which a reader can understand a (written) text. Automatic approaches to readability, also referred to as readability formulas, calculate a score based on measures such as word length, sentence length, and (difficult) word frequency. Many variations of these formulas exist, but the Flesch-Kincaid reading ease and Gunning-FOG index have become the most widely used, tested and reliable readability metrics (Zhou et al., 2017).

The Flesch-Kincaid reading ease (Kincaid et al., 1975) formula is defined as:

$$ReadingEase = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$$

The Gunning-FOG (Gunning, 1952) is defined as:

$$Gunning - FOG = 0.4\left[\left(\frac{words}{sentences}\right) + 100\left(\frac{complex\ words}{words}\right)\right]$$

A limitation of the Gunning-FOG index is that it defines complex words as words with at least four syllables. This means that words such as 'interesting' are classified as a complex word, while it probably will not increase the difficulty a person has when reading a text that contains this word.

Lexical diversity is defined as the variety of words used in a text. Lexical diversity indices generally measure the number of unique words occurring in the text relative to the text's total number of words. In the field of linguistic research, lexical diversity indices have been found to be significant predictors of human evaluation of text quality (McCarthy & Jarvis, 2010). Since most lexical diversity indices are based on the ratio of unique words to the total number of words in a text, they have been found to be sensitive to variations in text length. McCarthy and Jarvis (2010) propose an alternative to the traditional lexical diversity indices with their measure of textual lexical diversity (MTLD) and show that their approach is not biased towards text length. Still, they advise researchers to consider using multiple lexical diversity indices in their studies rather than any single index alone, as each index captures unique lexical information. Based on their findings, HD-D, and MLTD are used to measure the lexical diversity of product descriptions. Hypergeometric distribution D (HD-D), developed by McCarthy and Jarvis (2007), calculates for each lexical type in a text the probability of encountering any of its tokens in a random sample of 42 words drawn from the text. The Measurement of lexical textual diversity (MLTD) index is defined as the mean length of word strings in a text that maintain a given TTR value.

To evaluate GPT-3's instruction methods on readability and lexical diversity, twenty product descriptions are generated (five male product descriptions with haptic information, five male product descriptions without haptic information, five female product descriptions with haptic information, five female product descriptions without haptic information) for each instruction method (zero-shot to finetune). To account for variance in the generated product descriptions, the mean readability and lexical diversity scores over each instruction method are reported along with confidence intervals.

4.4.2 Manual Error Analysis

Opposed to automatic evaluation and human evaluation, reporting the errors in NLG outputs provides meaningful insight into the strengths and weaknesses of NLG systems. However, in the field of NLG, a severe under-reporting of NLG error analysis exists, as most papers neither include any error analysis nor provide any examples of errors made by the system (Van Miltenburg et al., 2021). The just-referenced authors highlight the necessity of manual error analysis, as it shows the complexity of the task at hand, and the challenges that still lie ahead. In addition to addressing the lack of reported error analysis, Van Miltenburg et al. (2021) provide recommendations in error identification and express the importance of both stating a clear goal of the error analysis and describing the expected errors of the NLG output. As described in See et al. (2017), seq2seq models, including GPT-3, commonly suffer from repeating already stated information. Moreover, they often hallucinate, which appears when there is no relation to generated words and the provided input sequence.

Therefore, identifying repetition and hallucinations is the main goal of the manual error analysis conducted in this paper. Since this study concerns data-to-text generation, a differentiation is made between hallucinations (generating words with no relation to the input data), and factual errors, which happen when the product description contains the wrong product attributes (for example black jeans instead of blue jeans). If additional errors are identified that do not correspond to these categories, new categories for these errors will be formed. Due to time and financial constraints, the product descriptions will be annotated by myself. This implies that the detection of errors may be prone to individual bias as I was involved in training and experimenting with the NLG system that generates the product descriptions. Overall, a top-down approach is used for the expected errors, while a bottom-up approach is used for unforeseen errors. For each error category, the total number of identified errors is reported, and examples of the errors are provided. For the manual error analysis, the same data is used as in the automatic evaluation, resulting in a total sample of 80 product descriptions.

4.4.3 Human Evaluation Method

Based on best practices in the human evaluation of NLG and inspired by Qader et al. (2018) and van der Lee et al. (2018), the to-be performed human evaluation aims to evaluate the generated product descriptions on readability, factual correctness, and unnecessary repetition, with the corresponding questions to these criteria shown in table 3.

Table 3

Criteria human evaluation of product descriptions

Criteria	Question
Readability	The text in this product description flows in a natural manner and is easy to read.
Correctness	The product description contains grammatically correct sentences.
Repetition	The product description does not contain unnecessary repeated information.

All scales are measured using a 7-point Likert scale (1=strongly disagree, 7=strongly agree), and order randomization is applied. Next to evaluating the quality of GPT-3-generated product descriptions, the human evaluation study also validates the haptic information method. By asking participants to rate the degree of haptic information in product descriptions, the logit bias method used in generating product descriptions with and without haptic information is validated. The two haptic criterion questions are shown in table 4.

Table 4

Haptic information questions

Criteria	Question
Haptic information	This product description contains words that describe how the product feels.
	This product description contains words that describe what touching the product would feel like.

To counter potential gender bias and help respondents identify with the products, male/female participants are shown a single male/female pair of jeans six times, joined with a generated product description with or without haptic information (three times each). As participants rate the product descriptions on five questions each time, they answer a total of 30 questions. The product descriptions selected are sampled from the 80 product descriptions generated for the automatic evaluation and manual error analysis. To minimize a possible bias towards the length of a product description, for each product description category (e.g. women's finetune product description with haptic information product description), the product description closest to the average length of all product descriptions is selected. An overview of the selected product descriptions and the product images are shown in tables 2 and 3 of appendix B and the full survey is shown in appendix C.

4.5 Consumer Purchase Intention Experiment Method

For the consumer purchase intention and AI distinguishability experiment, a 2×2 within-subjects experiment was run. Participants were shown all experimental conditions in random orders: (1) GPT-3-generated product descriptions, with non-haptic information; (2) GPT-3-generated product descriptions with haptic information; (3) human-written product descriptions with non-haptic information, and (4) human-written product descriptions with haptic information. The participants were recruited by means of convenience sampling of colleagues and friends and the experiment was conducted with the online survey platform Qualtrics.

To test NFT's moderating effect and measure the individual difference in NFT, the scale developed by Peck and Childers (2003) was used. Their developed NFT scale exists of an autotelic and an instrumental dimension. The autotelic dimension refers to touch as a hedonic-oriented response seeking fun, arousal, and enjoyment, while the instrumental dimension refers to outcome-directed touch with a purchase goal. The scale consists of six instrumental and six autotelic statements shown in table 5. Participants indicated to what extent they agree with the statements on a Likert scale ranging from strongly disagree (1) to strongly agree (7). To prevent information overload resulting in possible discontinuation of participation, participants were first shown a matrix form with the instrumental statements and then the autotelic statements.

Table 5

The two dimensions of NFT and the scale items.

	Instrumental		Autotelic
1	I place more trust in products that can be touched before purchase.	1	When walking through stores, I can't help touching all kinds of products.
2	I feel more comfortable purchasing a product after physically examining it.	2	Touch products can be fun.
3	If I can't touch a product in the store, I'm reluctant to purchase the product.	3	When browsing in stores, it is important for me to handle all kinds of products.
4	I feel more confident making a purchase after touching a product.	4	I like to touch products even if I have no intention of buying them.
5	The only way to make sure a product is worth buying is to actually touch it.	5	When browsing in stores, I like to touch lots of products.
6	There are many products that I would only buy if I could handle them before purchase.	6	I find myself touching all kinds of products in stores.

Note. from "Individual differences in haptic information processing: the 'need for touch' scale.", by Peck and Childers (2003), Journal of Consumer Research, Volume 30 Issue 3, p432.

Since the experimental setup of the purchase intention experiment in this study is inspired by the work of Silva et al. (2021), the generated product descriptions were designed to follow a similar structure. Their product description used in the experiment states:

"*Sportive hoodie, slim fit. Soft, pleasant touch, ideal for cooler days. The lining has a velvety, soft feel. Fleece hood, metal zipper, and two kanga pockets at the front. White polyester hood cord. Blended fabric (65% cotton, 35% polyester*). (Silva et al., 2021, procedure and stimuli section, para. 2)"

This means that, consistent with Silva et al. (2021), the generated product descriptions are intended to start with the product, followed by the main attributes, and ending with the material composition.

Silva et al. (2021) claim they did not find a moderating relationship for NFT due to the simplicity of the product, a grey hoodie, in their experiment. To mitigate the risk that a too simplistic product influences the results, this study shows two products to participants for each experimental condition (haptic and non-haptic information): a presumably simplistic product and a presumably more sophisticated product, in which fabric and touch play a more important role. Moreover, to prevent a varying stimulus effect of the product images between male and female participants, the simplistic and sophisticated product shown to the male and female participants are constrained to be equal. This means that a dress, for example, cannot be included in the experiment, as this product exclusively targets women. Following Silva et al. (2021), a hoodie is used as a simplistic product. For the sophisticated product, jeans are used, since the main attributes of jeans refer to the texture, fit, and fabric.

For the product images and the human-written product descriptions, the product images and descriptions from the online fashion e-commerce retailer Nordstrom were used. Since Nordstrom is one of the largest fashion retailers in the United States, the assumption is made that their product descriptions are (human-)written by expert copywriters (Statista, 2021). However, unlike the structure in the study of Silva et al. (2021), the product descriptions of Nordstrom only contain information about the product category and the product's attributes, and the products materials are listed in bullet points. Therefore, these bullet points are rewritten into sentence structure and appended to the Nordstrom product descriptions. Additionally, since the product descriptions of Nordstrom naturally focus on the haptic attributes and information of a product, various product descriptions are rewritten into non-haptic product descriptions by rewriting touch-related phrases into more general phrases. This results in the product descriptions shown in table 6.

Table 6

Human-written product descriptions

Women		**Men**	
Hoodies	**Jeans**	**Hoodies**	**Jeans**
A slouchy oversized fit lends a borrowed-from-him look to a cozy pullover hoodie made from an eco-friendly recycled -cotton blend. The hoodie features a kangaroo pocket and it is made from 80% organic cotton, and 20% polyester.	Rock a modern casual-cool silhouette in these high-waist jeans punctuated with elongating wide legs. They are made from 99% cotton and 1% spandex, and the medium dark indigo wash creates a wear-with-everything versatility!	Crafted from an organic cotton blend, this hoodie is ready for the weekend. The hoodie has drawstring hood and a kangaroo pocket, making it a wardrobe essential! It is made from 80% organic cotton, and 20% polyester.	Sanding at the puckers and wear points burnish the faded wash of jeans cut slim with a modern straight-leg profile that looks great day or night. The jeans have a zip fly with button closure, and they are made from 51% rayon, 26% cotton, 22% polyester, and 1% spandex, to provide you with everyday comfort!
Layer up with comfort at the gym, at home or while running errands in a classic cotton blend hoodie styled with logo-embroidered lettering at the chest. The hoodie is made from a 65% cotton and 35% polyester blend, which keeps it looking good and feeling soft for a long time.	New Roadtripper Authentic denim brings an old-school, rigid look and tons of stretch to these comfy skinny jeans reengineered for those with an hourglass shape. Made from a blend of 76% cotton, 22% recycled polyester, and 2% spandex, these jeans provide a comfortable yet stylish fit.	This easy hoodie knit from a soft cotton blend is an essential layering piece for the daily rotation. It features a kangaroo pocket and is made from 75% cotton and 25% polyester.	Sanded and faded to old-favourite perfection, these contour-hugging jeans sport plenty of stretch and the enhanced comfort of sustainably produced cotton. This five-pocket style jeans has a zip fly with button closure, and is made from 95% cotton, 4% elastomultiester, and 1% elastane which ensures shape preservation.

Due to the randomness in the GPT-3-generated product descriptions, which could lead to errors identified in the manual error analysis, consistently generating flawless product descriptions for the experiment was considered impossible. Consequently, testing whether participants can distinguish between randomly selected product descriptions and human-written is trivial, as participants will easily distinguish the GPT-3-generated product descriptions due to the errors. Therefore, a human-in-the-loop method is used to select the highest quality product descriptions. For each product, three product descriptions are generated with finetune instruction and the product description with the fewest mistakes is selected. Moreover, the deviation to the mean length of the human product descriptions is considered by only selecting descriptions that deviate at most six words from the mean length. This method results in the eight product descriptions shown in table 7.

Table 7

GPT-3 generated product descriptions

Women		**Men**	
Hoodies	**Jeans**	**Hoodies**	**Jeans**
A slouchy fit brings soft structure to a versatile cotton blend hoodie with a bit of stretch for shape and comfort. The hoodie has a drawstring hood and long sleeves with ribbed cuffs. It's made from 64% cotton and 36% polyester for comfort wearability in a variety of situations!	A slim leg achieves a sleek profile in these cropped straight jeans designed with a high rise waist. This jeans has a zip fly with a button closure and a Five pocket design. The jeans are made from 99% cotton and 1% elastane for comfort during wearability!	A soft cotton blend drawstring hoodie with a smooth finish and curved hemline is a comfortable way to warm up on a chilly day. The hoodie has a drawstring hood and ribbed cuffs and hem. It's made from 80% cotton and 20% polyester for comfort wearability in any season!	A slim leg achieves a clean profile in these smooth, dark-washed skinny jeans that are made for every body type. The jeans has a button fly and comes in a five pocket style. It's made with 94% cotton, 4% polyester and 2% elastane for a comfortable fit that will last!
A slouchy kangaroo pocket covers this cozy, drawstring hoodie that's ideal for relaxing at home or heading to the gym. The hoodie has drawstring hood and comes in a pullover style. The fabric is 50% polyester, 46% cotton, 4% rayon for comfort in any	Soft light-blue wash furthers the casual, comfortable feel of these slim-tailored jeans cut from Italian denim that's so comfortable you'll never want to take them off. This jeans has a Zip fly with button closure and it comes in a Five pocket style. It's made from 52% cotton, 34% lyocell, 12% polyester, and 2% spandex.	With drop sleeves and loose-fitting profile, this fleece hoodie will keep you cozy and dry in stormy conditions or while you're hanging with friends. The hoodie has a drawstring hood and kangaroo pocket. It's made from 100% cotton for a comfortable fit!	A classic for casual days, these faded blue jeans have a mid-rise waist and are designed with a bit of stretch for versatility. The jeans has a Zip fly with button closure and Five pocket style. They're made from 98% elastane and 2% cotton to provide comfort during wearability.

For each product / product description pair, participants indicated their purchase intention by expressing their overall feeling about the product, and their intend to purchase on a 7-point Likert scale. To indicate whether participants think a product description is generated by AI or human-written, participants select one of the options. Table 8 shows the questions asked to participants.

Table 8

Purchase Intention Experiment Questions

Concept	**Question**	**Scale**
Purchase Intention 1	If you needed a new hoodie, would you order this product to try it out?	1 = Extremely unlikely 7 = Extremely likely
Purchase Intention 2	Please describe your overall feeling about the product.	1 = Definitely do not intend to buy. 7 = Definitely intend to buy.
Author	The product description was…	1. Written by a human 2. Generated by AI

To minimize order effects, order randomization was applied to the order in which products image and description pairs are shown. After indicating the questions on the NFT scale, participants were shown an instruction page on which they are asked to imagine that they are surfing the Internet looking to buy new clothing, specifically a hoodie and a new pair of jeans. Next, they were informed that their task was to express their overall feelings about the products displayed and to indicate whether they think a product description is AI-generated or human-written. A full overview of the survey is shown in appendix D.

6. Results

This section describes the results of the various methods used to test the eight hypotheses. The first the subsection reports the results of CLIP in category prediction and feature extraction. The second subsection discusses the results of the human evaluation, automatic evaluation, and manual error analysis performed to evaluate the quality of the generated product descriptions. The third subsection discusses the results of the purchase intention consumer experiment. By reporting the results, it is determined which hypothesis are supported and which are not.

5.1 CLIP Results

To determine which out of the six available model variations of CLIP best suited the apparel product classification and attribute extraction task, all available models were tested on the DeepFashion category classification task. The performance of the models is shown in table 6.

Table 8

the performance of different CLIP models on category prediction

Model	Top 1	Top 3	Top 5
RN50	0.2727	0.4727	0.6
RN101	0.2909	0.5595	0.7131
RN50x4	0.3495	0.5838	0.6969
RN50x16	0.4	0.6384	0.7434
ViT-B/32	0.3656	0.5797	0.6848
ViT-B/16	0.3738	0.6364	0.7494

Overall, the ResNet50x16 and ViT-B/16 models obtained the best performance on DeepFashion. Interestingly, the ResNet50x16 model obtained the best top-1 and top-3 accuracy, while the ViT-B/16 obtained the highest top-5 accuracy. Since the ResNet50x16 model scored best on top-1 accuracy, this architecture is used to evaluate the performance of CLIP in this study.

Ideally, CLIP would be compared against a fully supervised model that holds the state-of-the-art performance on the DeepFashion dataset. Unfortunately, confirming the replication problem in the field of deep learning, none of the state-of-the-art papers on this dataset have published the code leading to their results (Huynh & Elhamifar, 2020; Zhang et al., 2020). Consequently, replicating the results of these papers was almost impossible. Since Radford et al. (2021) compared the performance of a zero-shot CLIP classifier with a fully supervised ResNet-50 model, this model was used to compare the performance of CLIP on the DeepFashion benchmark. The training procedure of the ResNet-50 model on DeepFashion was based on a GitHub repository training a ResNet-34 model on DeepFashion (Sennikova, 2021).

Inspired by constructs in the cerebral cortex of the brain, the ResNet architecture introduced gated shortcut connections in deep neural networks (He et al., 2016a). The core idea of a ResNet is the identity shortcut connection, which avoids the vanishing gradient problem commonly observed in a deep neural network. As a result, and contrary to earlier deep neural networks, the performance of a ResNet no longer saturates as more layers are added, which means that the ResNet is equally capable of detecting low-, mid-, and high-level features in images. The ResNet architecture, as introduced by He et al., (2016a) consisted of 34 layers. The ResNet-50 model is a modified design of the ResNet-34 model and uses stacks of 3 layers instead of 2, resulting in a total of 50 layers and achieving much higher accuracy (He et al., 2016b).

Table 9 shows the product classification and attribute extraction performance of a ‘regular’ CLIP and a prompt-engineered CLIP compared to the trained ResNet-50 model and the non-replicated results of the FashionNet model, the best performing model in the initial DeepFashion publication, as reported in Liu et al. (2016). CLIP’s zero-shot top-1 accuracy on category prediction is 0.36 and increases to 0.579 and 0.68 for top-3 and top-5 accuracy. In comparison, a supervised ResNet-50 model achieves 0.6885 top-1 accuracy, equalling the top-5 accuracy of regular CLIP. The top-5 accuracy of the ResNet-50 model is quite impressive, as the correct category is included in the top-5 predictions in 92% of the instances. This exceeds the performance of FashionNet, the best performing model in the DeepFashion paper, which achieved a 0.9 top-5 accuracy.

Table 9

DeepFashion performance of task-specific model vs CLIP

	Product Category			Texture		Fabric		Shape		Part		Style	
Model	**top-1**	**top-3**	**top-5**	**top-3**	**top-5**	**top-3**	**top-5**	**top-3**	**top-5**	**top-3**	**top-5**	**top-3**	**top-5**
FashionNet (not replicated)	-	0.8258	0.9017	0.3746	0.4952	0.3930	0.4984	0.3947	0.4859	0.4413	0.4413	0.6643	0.7316
ResNet-50 (Replicated)	0.6728	0.8650	0.9276	-	-	-	-	-	-	-	-	-	-
regular CLIP	0.365	0.5797	0.68	0.0191	0.0301	0.0143	0.0241	0.0152	0.0295	0.0327	0.0437	0.1420	0.1683
Prompt-engineered CLIP	0.4263	0.6808	0.7920	0.0837	0.1223	0.0665	0.0882	0.0915	0.1453	0.0783	0.1220	0.1079	0.1390

Unfortunately, this study failed to train a ResNet-50 model on the attribute extraction task, which means the only reference point is the non-replicated reported results of the FashionNet model in Liu et al. (2016). Although the FashionNet model originated in 2015, it still significantly outperforms CLIP on attribute prediction, with recall rates ranging from a fivefold to a tenfold of a regular CLIP. On attribute extraction, a regular CLIP hardly achieves any performance, as the top-3 recall is close to zero in some attribute categories. Prompt engineering results in a significant increase in performance, as it increases the top-k product category accuracy by 9% on average. Similarly, for attribute prediction, prompt engineering leads to a significant increase in recall. Although the recall is still a fraction of the FashionNet model's, prompt engineering doubles the recall on average and in some categories even reaches a three- or fourfold compared to a regular CLIP. A slight decrease is only observed in the attribute category style, which is quite remarkable. Possibly, the words in this attribute category have a more stand-alone meaning and do not require additional context to increase the model's understanding of these words. Embedding them with the additional product category could then maybe lead to the embedding being more skewed to the clothing category instead of the attribute, which results in worse embeddings compared to non-engineered embeddings.

Looking for an answer that explains why CLIP is especially underperforming on attribute extraction, several product images are inspected on their attributes. A sample of the inspected product images is shown in figure 1 in appendix A. Generally, the attributes of a product only contribute to a minor area of the product image. Since CLIP is a visual language model which compares the textual embedding to the image embedding, the image embedding is likely embedded as the total image (for example ‘a woman with a red shirt and black jeans’ in the left top of figure 1 in appendix A). Hence, subtle attributes may not have been included in the data on which CLIP was trained, meaning the model is unlikely to include these subtle attributes in the textual embedding of the image. Moreover, some attributes are simply very hard to detect. An example of this is given on the top right corner of figure 1 in appendix A. The attribute ‘cargo’ refers to the fact that the pants have two additional pockets on the side of the trouser leg. In this specific image, due to the light colour and bag held beside the leg, it is quite hard to detect that these trousers are in fact cargo trousers.

Because CLIP calculates the distance between an image embedding and a textual embedding, it encounters the problem of polysemy. In Radford et al. (2021, p. 7) polysemy is described as follows: “When the name of a class is the only information provided to CLIP’s text encoder it is unable to differentiate which word sense is meant due to the lack of context.” An example of a label sensitive to polysemy is the case of the attribute ‘cargo’, as the word refers to both a type of jeans and goods carried on a ship, aircraft, or motor vehicle. If the textual encoder embeds the word referring to the latter definition, it logically will not detect this attribute, because many of the other attributes which have been embedded as an attribute related to clothing will automatically have a mathematically closer embedding to the image embedding.

Concludingly, both a standard CLIP and a prompt-engineered CLIP achieve inferior accuracy and recall on the product classification task and the attribute extraction task compared to the ResNet-50 and FashionNet model. Consequently, **H1** is not supported.

5.2 Product Description Evaluation Results

5.2.1 Automatic Evaluation Results

Table 10 shows the scores for the readability formulas and lexical diversity indices. The scores are the average of the scores calculated on the twenty generated men's and women's clothing product descriptions (ten for each gender). Within the ten generated men's and women's clothing product descriptions, five product descriptions were generated with haptic information and five product descriptions were generated without haptic information. The non-averaged readability and lexical diversity scores of the generated men's and women's clothing product descriptions separately as well as an overview of the generated product descriptions can be found in table 4, 5, 6, and 7 in appendix B.

Table 10

Readability and Lexical Diversity Scores

Method	Flesch	Gunning	MTLD	HD-D
Zero-shot	70.322	7.8215	39.4915	0.2163
	(62.2472, 78.3968)	(5.8498, 9.7932)	(30.1395, 48.8435)	(0.07, 0.3627)
One-shot	67.535	7.251	70.6711	0.1134
	(56.0974, 78.9726)	(6.1087, 8.399)	(48.6597, 92.6825)	(-0.0163, 0.2432)
Few-shot	73.043	8.7485	78.4371	0.4223
	(68.1933, 77.8927)	(7.2747, 10.2223)	(57.636, 99.2382)	(0.2192, 0.6253)
Finetune	76.6075	8.02	82.5203	0.7562
	(73.7288, 79.4862)	(7.2913, 8.7487)	(74.392, 90.6487)	(0.6339, 0.8786)

Note. 95% confidence intervals in parenthesis

On readability, the instruction methods score differently on the Flesch-Kincaid reading ease and the Gunning-FOG index scores. On Flesch-Kincaid reading ease, the product descriptions become more easily readable as the product description generation is based on more data; the readability increases by 7 points on average on the Flesch-Kincaid reading ease scale from zero-shot to finetune instruction. However, this difference is not statistically significant, as the 95% confidence interval of the zero-shot and finetune method Flesch-Kincaid scores overlap each other. The Gunning-FOG index shows a different pattern, as zero-shot and few-shot instruction methods score higher than one-shot and finetune methods. As the Gunning-FOG index calculates the ratio of the number of words to the number of sentences and the ratio of complex words to total words, this means that on average, product descriptions generated by few-shot and finetune methods contain either relatively more complex words or relatively longer sentences compared to product descriptions generated by zero-shot and one-shot methods. On both the Flesch-Kincaid grade level and the Gunning-FOG index, the confidence intervals of the finetune-generated product description are considerably more narrow compared to the confidence interval of zero-shot-generated product descriptions. This shows that the finetune instruction method generates more stable product descriptions concerning readability.

On lexical diversity, the results indicate more consistent and statistically significant findings. On both MTLD and HD-D, the few-shot and finetune methods generate statistically significantly more diverse product descriptions compared to those generated by a zero-shot method. Concludingly, the only statistically supported claim that can be made is that product descriptions generated by few-shot and finetune methods have a higher lexical diversity compared to those generated by a zero-shot method. Due to the high variation in readability scores for product descriptions generated by the same instruction method, resulting in wide confidence intervals, no statistically supported claim concerning readability can be made. This means that **H2** is not supported: no statistical evidence is found that few-shot and finetune-generated product descriptions are more readable than zero-shot and one-shot generated product descriptions. Moreover, only zero-shot-generated product descriptions are found to be less lexically complex than few-shot and finetune-generated product descriptions, while one-shot-generated product descriptions have approximately the same lexical diversity.

5.2.2 Manual Error Analysis Results

By annotating the eighty generated product descriptions shown in tables 4 and 5 in appendix B, errors were identified and categorized into four main categories. After examining the product descriptions, it was decided to distinguish between two types of hallucinations. Errors categorized as task hallucination refer to generated product descriptions in which GPT-3 did not comprehend the task of product description generation, resulting in unrelated text generation. Errors categorized as product hallucinations refer to generated product descriptions in which GPT-3 included certain features that are not present in the input data. Moreover, from a bottom-up method, errors related to copying the input data and unfinished sentences were identified. The copying of the input data error is defined as the generated product description being (almost) completely identical to the input prompt. The unfinished sentence error is defined as product descriptions in which the final sentence was not finished correctly (with a period or exclamation mark). Repetition errors refer to product descriptions in which the same information was provided more than once. Errors categorized as factual incorrectness are defined as product descriptions which contained incorrect descriptions of attributes.

Table 11

Errors per error category for each instruction method.

Method	Copy of data input	Unfinished sentences	Repetition	Factual incorrectness	Product hallucination	Task hallucination	Total method mistakes
Zero-shot	5	18	4		3	6	36
One-shot		6	1	3	11	1	22
Few-shot		6		2	5		13
Finetune		2		4	11	1	18
Total category mistakes	5	32	5	9	30	8	93

Note. A single product description can contain multiple mistakes; hence the total number of mistakes is larger than the total number of product descriptions annotated.

Table 11 shows the number of errors in each error category for the instruction methods. Unfinished sentences and product hallucinations were the most identified errors, followed by factual incorrectness and task hallucination. The zero-shot-generated product descriptions contained the most errors, while few-shot-generated product descriptions contained the fewest. Unfinished sentences, task hallucination, repetition, and copy of data input errors occurred predominantly in the zero-shot-generated product descriptions, while factual incorrectness and product hallucination mostly occurred among one-shot, few-shot, and finetune-generated product descriptions. The vast occurrence of task hallucination errors in zero-shot-generated product descriptions shows that, without a provided example, GPT-3 occasionally fails to understand the requested task to be performed. Moreover, not providing any examples for the product descriptions generation task frequently led to a near-exact copy of the input data being returned as output text.

An overview of typical examples of the identified errors is shown in table 8 of appendix B, with identified error highlighted in red in each sentence. While this table aims to show an overview of the identified errors, errors placed in the same category differ significantly. For example, errors categorized as product hallucination error contain extraneous unverified information of all sorts such as production country, welt pocket, waistband, distressed knees, material, and rise. Hallucinations regarding the jeans type (low-, mid-, or high-rise) were the most prevalent, occurring eighteen times. Problematically, GPT-3 did not hallucinate consistently but described the same jeans as low-, mid-, or high-rise on multiple occasions. Hallucinating the attribute waist type occurred twelve out of eighteen times in few-shot and finetune-generated product descriptions. An explanation for this could be the fact that one of the examples in the few-shot prompts mentioned the waist type, and three examples in the finetune training data contained the waist type. As the instruction method highly influences the output GPT-3 generates, containing these words in the prompt or training data will probably cause an increased probability of sampling these words in the product description generation.

In the category factual incorrectness, the most often occurring error concerned the wrong type of fit (skinny, slim-fit, straight-leg) being mentioned, or the product description containing a contradicting statement and mentioning both skinny and straight-leg. These errors were categorized as factual incorrectness, because in contrast to the hallucination errors, the input data stated the jeans were skinny fit. Like the hallucination errors, the origin of factual inaccuracies in fit type probably is the prompt instruction. The single example provided in the one-shot instruction method concerns a straight-leg jeans and this example is also used in the few-shot prompt, while the finetune training data contains three examples of a straight-leg jeans. Although the input sequence mentions that it concerns a skinny-fit jeans, the examples provided in the prompt about a straight-fit jeans could influence GPT-3 into sampling this feature information in the generated product description. Moreover, factual inaccuracies in which two fit types were included in the product description show that GPT-3 is incapable of reasoning that a pair of jeans cannot be a skinny fit and a straight fit at the same time.

The most frequently occurring error is unfinished sentence, which occurs in 36 out of eighty product sentences. This error is most probably a result of the max tokens parameter in the prompt, as this limits the product descriptions to a certain length. Although these errors can be easily solved with post-processing and removing unfinished sentences, this would result in omitted product information. Moreover, one would expect that when GPT-3 nears a token limit, it returns a text in which the final sentence is completed instead of cutting off the sentence mid-word.

To summarise, regardless of the way in which GPT-3 is instructed, like other generative language models, GPT-3 still suffers from hallucination. However, as GPT-3 is instructed with more examples, the errors shift from total task incomprehension or returning a copy of the input prompt to smaller context-specific hallucinations and factual inaccuracies. These factual inaccuracies and hallucinations in, for example, fit and waist type show the importance of careful prompt design for GPT-3 to better leverage GPT-3's single-shot, few-shot, and finetune capabilities. Taking the errors identified in the zero-shot-generated product descriptions, the average readability and lexical complexity scores, and the time constraints of the human evaluation study into account, it is decided to omit the zero-shot generated product descriptions from the human evaluation study. Consequently, only one-shot, few-shot, and finetune generated product descriptions are evaluated by human participants on readability, correctness, and repeated information.

5.2.3 Human Evaluation Results

For the human evaluation experiment on product description quality, a Qualtrics survey was distributed among a convenience sample of friends and colleagues. In total, 24 participants completed the survey, of which seventeen were male and six were female. Table 12 reports the average ratings of the three criteria on which participants rated the product description.

Table 12

Mean human evaluation ratings per criteria

	Readability	**Correctness**	**Repetition**
One-shot	3.8696	4.5435	2.6740
Few-shot	4.9348	5.4783	3.1740
Finetune	5.4565	4.6087	2.2340

To test whether statistically significant differences in mean readability, correctness, and repetition score exist between the instruction methods, a one-way analysis of variance (ANOVA) test is performed on each criterion. The results show that a significant difference exists between instruction methods on readability (F (3, 20) = 11.7611, p = 1.956e-05), correctness (F(3, 20) = 5.8355, p = 0.0037), and unnecessary repetition (F(3, 20) = 5.0497, p = 0.0076). Since this indicates significant differences in mean readability, correctness, and repetition between the instruction methods, several t-tests were performed to indicate which differences between the instruction methods were statistically significant.

On readability, the difference between few-shot (M = 4.9348; SD =1.5550) and one-shot (M = 3.8695; SD = 1.9277) was significant (t (22) = 2.9171; p < 0.0045), the difference between finetune (M = 5.4565; SD = 1.2420) and one-shot (M = 3.8695; SD = 1.9277) was significant (t (22) = 4.6936; p < 1.15e-05), but the difference between few-shot (M = 4.9348; SD =1.5550) and one-shot (M = 5.4565; SD = 1.2420) was not significant (t (22) = 1.7781; p < 0.0789) at a confidence level of 95%. This means few-shot and finetune-generated product descriptions are perceived as more readable compared to one-shot-generated product descriptions.

On factual correctness, the difference between few-shot (*M* = 5.4782; *SD* = 0.9600) and one-shot (*M* = 4.5434; *SD* = 1.6015) was significant (t (22) = 3.3954; $p < 0.0011$), the difference between finetune (*M* = 4.6087, *SD* = 1.7189) and few-shot (*M* = 5.4782; *SD* = 0.9600) was significant (t (20) = 2.9955; $p < 0.0038$), but the difference between finetune (*M* = 4.6087, *SD* = 1.7189) and one-shot (*M* = 4.5434; *SD* = 1.6015) was not significant (t (20) = -0.1883; $p < 0.8510$). Therefore, few-shot generated product descriptions are perceived to contain significantly more factually correct information compared to one-shot and finetune generated product descriptions.

On unnecessary repetition, the difference between few-shot (*M* = 3.1739; *SD* = 1.7423) and one-shot (*M* = 2.6739; *SD* = 1.3833) was not significant (t (22) = 1.5243; $p < 0.1311$), the difference between finetune (*M* = 2.2391; *SD* = 1.0151) and one-shot (*M* = 2.6739; *SD* = 1.3833) was not significant (t (22) = -1.7186; $p < 0.0894$), but the difference between finetune (*M* = 2.2391; *SD* = 1.0151) and few-shot (*M* = 3.1739; *SD* =1.7423) was significant (t (20) = 3.14409; $p < 0.0024$). Consequently, few-shot generated product descriptions are perceived as significantly more repetitive compared to finetune generated product descriptions.

Since male participants were shown male product images and product descriptions and vice versa, a post-hoc analysis was performed to validate whether results were consistent between the male and female participants. Table 10 in appendix B reports the mean ratings of the three criteria filtered on gender. The performance of isolated t-tests for readability, correctness, and repetition scores between male and female participants was problematic given the small sample size of female participants. This meant that only the difference on unnecessary repetition between few-shot and finetune instruction was consistently significant in both the male participant and female participant group. Although no statistically significant contradicting findings were found between the male and female participant group, this was mainly caused by the statistical insignificance of the eight other t-test comparisons in the female participant group. The mean readability, correctness, and repetition scores between the two groups, shown in table 10 of appendix B, indicate several contradictions. For example, the one-shot product description shown to female participants was perceived as more readable than the few-shot product description while the male participants indicated the opposite. Moreover, the male participants perceived the few-shot product description to contain the most factually correct information, while the female participants rated the few-shot and finetune product description as equally factually correct.

The inconsistent results are not just the result of different sample sizes, but also caused by the random nature and unpredictability of NLG models such as GPT-3. This is a common issue for studies seeking to explore the capabilities of NLG models as the unpredictability leads to less experimental control (Rahwan et al., 2019). While this study aimed to generate product descriptions with consistent quality for male and female products, the human evaluation results show that the different attributes combined with the randomness of the model result in product descriptions with a disparity in quality.

Concludingly, **H3** is partly supported because although all statistically justifiable differences show a superiority of few-shot and finetune method generated product descriptions over one-shot generated product descriptions, the differences are not consistent between male and female participants. In addition, a statistically significant judgment cannot be made on all three criteria. While One-shot generated product descriptions are significantly less readable than few-shot and finetune generated product descriptions and one-shot generated product descriptions contain significantly less factually correct information than few-shot generated product descriptions, no statistically supported conclusion regarding one-shot versus few-shot and finetune can be made on unnecessary repetition.

Concerning the haptic method validation, the haptic product descriptions were perceived as containing significantly more haptic information compared to the non-haptic product descriptions. On the first haptic criterion (This product description contains words that describe how the product feels), the difference between non-haptic product descriptions (M = 3.6086; SD =1.6289) and haptic product descriptions (M = 4.7536; SD = 1.7440) was significant (t (22) = 3.9853; $p < 0.0001$). On the second haptic criterion (This product description contains words that describe what touching the product would feel like), the difference between non-haptic product descriptions (M = 3.5072, SD = 1.7288) and haptic product descriptions (M = 4.3623; SD = 1.715) was significant (t (22) =2.9170; $p < 0.0041$) as well. This means that **H4** is supported.

The Cronbach's alpha for the two questions on haptic information was .83. When asking participants whether they ran into any issues in the survey, someone mentioned that he intuitively assumed the two questions focused on different concepts instead of the same. As a result, this person gave different ratings for the two questions. In other words, although measuring a criterion with multiple 7-point Likert scales enabled multiple points of discrimination, it may also cause confusion among participants.

To summarise, the findings of the automatic evaluation, human evaluation experiment, and manual error analysis of the generated product descriptions do not unanimously indicate a single best GPT-3 instruction method. Although it is evident that few-shot and finetune instruction generate higher quality product descriptions compared to one-shot instruction, both methods have their advantages and disadvantages. While few-shot instruction results in more factually correct product descriptions, they contain more unnecessary repetitions compared to finetune generated product descriptions. Moreover, while few-shot generated product descriptions have fewer product hallucinations than finetune generated product descriptions, they also have more unfinished sentences. However, whereas product hallucinations such as 'low-rise' are unlikely to lead to suspicion that a product description is computer-generated, as it is difficult for a consumer to check, an unfinished sentence is more likely to raise distrust of the source. Besides that, the cost of GPT-3 finetune-generated descriptions are lower compared to few-shot-generated, because no examples need to be provided in the prompt. Consequently, although there are only minor differences in quality between few-shot and finetune generation, the practical benefits and reduction of 'easy AI signs' lead to the decision to use the finetune instruction method for generating product descriptions for the experiment in which participants try to distinguish GPT-3-generated product descriptions from human-written product descriptions.

To validate that finetune generated product descriptions with haptic information are perceived as more haptic compared to finetune generated product descriptions without haptic information, a T-test is performed. Like the general findings, the difference between non-haptic finetune generated product descriptions (M = 3.5217; SD = 1.6479) and haptic product descriptions (M = 5.6956; SD = 1.1845) was significant (t (22) = 5.1372; p < 6.11e-06) on the first haptic criterion. On the second haptic criterion, the difference between non-haptic product descriptions (M = 3.2608, SD = 1.6297) and haptic product descriptions (M = 5.1304; SD = 1.4864) was significant (t (22) =4.0647; p < 0.0002) as well.

5.3 Purchase intention experiment results.

For the purchase intention and AI distinguishability experiment, a Qualtrics survey was distributed among a convenience sample of friends and colleagues. In total, 48 participants filled in the survey, of which 25 were male, 21 were female, and 2 were unknown. Nine responses were not included in the analysis due to incompleteness or filling in the survey in less than a minute, resulting in a sample consisting of 39 responses.

The 12-item NFT scale, developed by Peck and Childers (2003), was used to assess individual differences in haptic information processing. The scale had an internal consistency of α =.93, with an internal consistency of α = .88 for the instrumental dimension and α = .94 for the autotelic dimension. The correlation between the instrumental and autotelic dimension was .61, which is in line with the findings of Peck and Childers (2003). A single NFT score per participant was determined by summing the twelve items and dividing by 12, which meant it ranged from 1 till 7. The average NFT of participants was 4.04, with a minimum of 1.16 and a maximum of 6.5. The participants purchase intentions were measured with two 7-point Likert scale items, which had an internal consistency of α = .92. The mean purchase intention was 4.176, which corresponds to 'Neither likely nor unlikely' and 'No preference' on the 7-point Likert scales.

To test the effect of haptic information on purchase intentions and the moderating effect NFT has on this relationship, an ordinary least squares regression (OLS) was performed on purchase intention with NFT and haptic information as independent variable, an interaction term between NFT and haptic information, and participants demographics (age and gender) as control variables. This regression provided satisfactory predictive power at F (8, 303) = 3.033, p = 0.0027, R = 0.074. As this model contained significant predictive power, the coefficients were inspected, which are shown in table 13.

As the table shows, neither haptic information (β = 0.20192, t = 0.116, p < 0.908) nor the interaction term between NFT and haptic information (β = 0.0059, t = 0.035, p < 0.0972) were significant. This means that both **H5** and **H6** are not supported. Remarkably, a significant positive relationship (β = 0.2369, t = 1.839, p < 0.067) for NFT is found, in contrast to what is suggested by the findings of Peck and Childers (2003) and Park (2006).

Table 13

OLS Regression coefficients

Predictor	b	SE	95% CI		p
			LL	UL	
Intercept	4.15	.16	3.83	4.47	<.001
Control Variables					
Age (25-34)	-.017	.23	-.63	.29	.47
Age (35-44)	-.40	.32	-1.02	.23	.21
Age (45-54)	-.70**	.29	-1.27	-.14	.02
Age (55-64)	.34	.54	-.72	1.40	.53
Gender (Female)	.35*	.19	-.03	.73	.07
Independent Variables					
Haptic Information	.019	.02	-.31	.35	.91
NFT	.24*	.13	-.02	.49	.07
Interaction term					
Haptic x NFT	.01	.17	-.32	.33	.97

Note. ****** p < 0.05, * p < 0.1, Age (dummy, reference category = Age (18-24)), Gender (dummy, reference category = Male), Haptic Information (dummy, 0 = without haptic information, 1 = with haptic information)

Besides NFT, gender and the age group (45-64) were found to be significant predictors of purchase intentions. In relation to the reference category Age (18-24), All the coefficients expect from the age group (55-64) are negative, which indicates that the products shown are primarily focused to address the fashion style of the reference age category. The fact that gender is a significant predictor of purchase intention indicates that female participants were more content with the products shown to them. Since Silva et al. (2021) attribute the failure to find a moderating relationship for NFT on the use of a hoodie as the garment in their experiment, it is tested whether the moderating relationship differs between hoodies and jeans. The product type appears to have no effect on the moderation effectiveness of NFT, since the coefficients of the interaction terms in the OLS regression of jeans and hoodies independently remain insignificant.

Concerning AI distinguishability, participants distinguished whether a product description was GPT-3-generated or human-written with an average accuracy of 48.48% (.95 percent confidence interval: 43.2, 53.77). A Wilcoxon signed-rank test is performed between the participants accuracies and a distribution which only consists of values of 0.5. The Wilcoxon signed-rank test indicates that the average accuracy does not significantly deviate from the 0.5 distribution (V = 162.5, p = .5077). In addition, a Bayesian binomial test (312 trials, 151 accurate guesses) yields a *BF-null* of 11.52 and a *BF-alt* of 0.087. Since the Bayes factor of the null hypothesis ('the guessing accuracy does not deviate from chance') is substantially higher than the Bayes factor of the alternative hypothesis ('the guessing accuracy deviates from chance'), it is concluded that the guessing accuracy does not deviate from random chance. In fact, a random guessing strategy would even outperform participants, given that participants are informed about the equal number of AI-generated and human-written product descriptions beforehand (which did not happen due to inducing an anchor effect). As a results, **H7** is supported. Moreover, controlling for age and gender revealed that no significant differences exist in the ability to detect the origin of product descriptions between demographical groups.

Despite not finding a positive relationship between haptic information in product descriptions and purchase intentions, it is still interesting whether participants respond similarly to AI-generated and human-written product descriptions. Therefore, the differences between purchase intentions on GPT-3-generated and human-written product descriptions were inspected, controlling on haptic information.

On the first purchase intention item (If you need a new hoodie, would you order this product to try it out?), the difference in purchase intention between human-written (M = 4.2051; SD = 1.607) and GPT-3-generated (M = 4.1667; SD = 1.6065) product descriptions with haptic information was insignificant (t (22) =-0.1550; p < 0.8770). Moreover, the difference in purchase intention between human-written (M = 4.1025, SD = 1.5) and GPT-3-generated (M = 4.2308; SD = 1.4590) product descriptions without haptic information was also insignificant (*t* (20) = 0.5412; $p < 0.5891$). Similarly consistent results were found on the second purchase intention item (Please describe your overall feeling about the product). Since the null hypothesis cannot that there is no difference in purchase intentions between GPT-3-generated product descriptions and human-written product descriptions cannot be rejected, **H8** is supported.

In light of the results reported in this section, table 14 provides an overview of the hypotheses formulated and what the reported results mean for the corresponding hypothesis.

Table 14

Summary of Results

Hypothesis	Variables	Result
H1	CLIP outperforms a supervised finetuned model on apparel category prediction and feature attraction.	Not supported
H2A	Fine-tune and few-shot GPT-3 generated product description score better on readability compared to zero-shot and one-shot GPT-3 generated product descriptions.	Not supported
H2B	Fine-tune and few-shot instructed GPT-3 generated product description score better on lexical diversity compared to zero-shot and one-shot instructed GPT-3 generated product descriptions.	Supported
H3	Participants evaluate fine-tune and few-shot instructed GPT-3 generated product description higher on readability, correctness, and repeated information compared to zero-shot and one-shot instructed GPT-3 generated product descriptions.	Partly Supported
H4	Participants observe that haptic product descriptions contain significantly more haptic information compared to non-haptic product descriptions.	Supported
H5	Participants experience a higher intention to purchase when viewing an apparel product with a haptic product description vis-à-vis an apparel product without a haptic product description.	Not supported
H6	NFT moderates the relationship between haptic information and purchase intention.	Not supported
H7	Participants are unable to distinguish GPT-3-generated product descriptions from human-written product descriptions.	Supported
H8	Participants express at least the same purchase intention when viewing a GPT3-generated apparel product description vis-à-vis a human-written product description.	Supported

7. Discussion

As previously mentioned, this study aimed to explore the generalizability of the zero-shot image model CLIP and the NLG model GPT-3 in fashion e-commerce to automatically provide a pleasant e-commerce experience by providing accurate and complete information product information in an appealing form. To this end, three research questions were formulated:

RQ1. Can CLIP accurately classify product categories and extract product features?

RQ2. How can the most accurate and appealing product descriptions be generated with GPT-3?

RQ3. Do GPT-3-generated product descriptions with haptic information increase consumer purchase intentions?

By comparing the performance of CLIP against finetuned ResNet-50 classifier on the DeepFashion benchmark, the inferiority of CLIP was shown. This means that at this moment in time, CLIP cannot accurately classify product categories and extract product features, as a considerably simple baseline outperforms it. Next, by conducting an automatic evaluation, a manual error analysis, and a human evaluation, the few-shot and finetune instruction methods proved to generate the most readable and correct product descriptions that contain noticeable haptic information, answering research question two. Finally, the purchase intention experiment results indicated that the human-written product descriptions with haptic information nor the GPT-3-generated product descriptions with haptic information resulted in increased purchase intentions.

Connecting these findings answers the main research question, which concludes that that zero-shot image and language models cannot be used to generate product descriptions with haptic information that increase consumers purchase intentions. Nevertheless, a number of interesting findings have resulted from this study, whose implications merit discussion.

One of the most remarkable results of this study is not finding support for the relationship between haptic information and purchase intentions and not finding support for a moderating relationship of NFT between haptic information and purchase intentions. A possible explanation for not finding support for these relationships may be that consumers still experience a need for touch, but it no longer influences their purchase decisions. Perhaps, due to the order-and-return culture of modern society and online shopping becoming the new standard, consumers have become accustomed to processing the product's physical attributes after purchase and returning a product if the physical attributes do not meet expectations. This culture change would mean that, contrary to Park (2006) and Silva et al. (2021), haptic information in product descriptions is no longer an effective way to compensate for not physically touching the products, as this inability is no longer perceived as a purchase barrier. Research on the validity of this idea is an essential step in establishing the role of need for touch in modern consumer behaviour, especially since Silva et al. (2021) do not find a moderating relationship for need for touch either.

Another central insight is that participants are incapable of distinguishing human-written product descriptions from GPT-3-generated product descriptions, given that a human is in the loop, selecting the best product description. With this finding, this study expands research into the distinguishability of AI-generated text (Gunser et al., 2021; Köbis & Mossink, 2021) to the field of consumer-oriented product descriptions. Although this demonstrates the progress in natural language programming, the indistinguishability of product texts can be misused to trigger unconscious buying behaviour. With the continuous research efforts of tech-giant funded AI research labs, it is not the question if but rather when more powerful language models will be released. Hence, it is inevitable that one day, natural language models that do not require human involvement will become the standard way of high-quality product description generation. Simultaneously, advanced analytics such as A/B tests will quickly indicate whether haptic information or other stimuli in product descriptions results in increased purchase intentions. Therefore, this study would like to emphasize that these methods should be used to provide a complete and user-friendly e-commerce experience and should not be used to 'hide' subconscious purchase triggers in product descriptions.

Although not connected to any hypothesis, the manual error analysis provided several valuable insights into the strengths and weaknesses of GPT-3, which is often forgotten in NLG research. The analysis showed that instructing GPT-3 with more examples reduces the total number of mistakes and shifts the type of mistakes made. Moreover, it demonstrated the randomness of GPT-3 and that a single perfect example is often counterbalanced by a multitude of product descriptions with minor errors such as factual incorrectness or product hallucination. The randomness and identified errors show that before models such as GPT-3 can be deployed without continuous scrutiny, significant research effort is required into reducing the randomness and enhance output consistency.

8. Limitations and Future work

Throughout this study, several limitations arose, which influenced the methods and results. This section discusses these limitations and the impact they have had on the execution of the study. Moreover, interesting ideas for future research are proposed, sometimes following from the limitations identified.

First, A considerable limitation in generating product descriptions with and without haptic information is the problem encountered with the logit bias parameter. The logit bias problem meant that longer touch-related words could not be encoded and prevented from being included in the generated product descriptions. This resulted in the inclusion of longer touch-related words such as 'supersoft' and 'comfort' in product descriptions intended to contain no haptic information. This problem was solved by manually rewriting sentences in the product description affected by this problem, which meant that the product descriptions without haptic information were no longer generated truly automatically. Moreover, due to the randomness in the GPT-3-generated product descriptions, it was decided to manually select the best product description out of three GPT-3-generated product descriptions. These two limitations made the product description generation method unintentionally result in a human-in-the-loop method.

While the explanation that NFT no longer influences purchase decisions provides interesting research opportunities, it would be too simplistic to pinpoint it as the sole culprit for not finding a direct effect for haptic information on purchase intention and not finding a moderating relationship for NFT. One limitation in the setup of the purchase intention experiment is that the decision to combine it with the GPT-3 distinguishability experiment, which resulted in less experimental control. This decision, combined with the decision to conduct a 2×2 within-subjects experiment, meant that participants had to be shown two different product images (e.g. hoodies) for the product description with haptic information and the product description without haptic information. These similar but different product images alongside the haptic and non-haptic product description may have induced unintentional varying image stimuli that trigger different purchase intention responses. Consequently, this could have negated the effect of haptic information.

In addition, the haptic information only accounts for several words in a product description. In for example the product description: 'A soft cotton blend drawstring hoodie with a smooth finish and curved hemline is a comfortable way to warm up on a chilly day. The hoodie has a drawstring hood and ribbed cuffs and hem. It's made from 80% cotton and 20% polyester for comfort wearability in any season!' only the words' soft', 'smooth', 'comfortable', and 'comfort' express haptic information. At the same time, the full product description consists of 49 words and covers other themes such as attributes and material composition. Research shows that when consumers devote their attention to an advertisement, they experience a reduced capacity to process all information they are exposed to and prioritize their attention to certain aspects of the presented text (Kong et al, 2019). Although the manipulation check in the human evaluation experiment demonstrated that the haptic product descriptions were perceived as containing significantly more haptic information, the haptic information may not have been visible enough. Consequently, participants may have devoted their attention to other aspects of the product descriptions or images, leading to a diminished haptic information effect.

Another limitation related to the purchase intention experiment is the assumption that product descriptions obtained from Nordstrom's website were human-written. This assumption was not verified, meaning that there is a chance that these product descriptions were in fact generated with the help of an NLG model. If this is the case, a logical consequence is that the accuracy with which participants distinguish GPT-3-generated product descriptions from AI-generated product descriptions from another source does not deviate from chance.

An additional research design decision that may have influenced the fact that participants could not distinguish the GPT-3 generated product descriptions is that the finetuned model was trained on similar product descriptions as those used in the purchase intention experiment. Because the finetuned model was trained on 100 product info - product description pairs of various product types in the assortment of Nordstrom’s online offering, the model likely has adjusted to the writing style and vocabulary used in the training data. Consequently, the GPT-3 is trained to generate product descriptions that look almost identical to the human-written product description obtained from Nordstrom, which logically complicates the distinguishability. On the contrary, the ability of GPT-3 to reflect the vocabulary of the dataset it is trained on while performing a specific task also creates interesting research directions. Solaiman and Dennison (2021) are one of the first to address this opportunity. They significantly change model behaviour by crafting and finetuning on a dataset to generate less abusive language. In the context of apparel e-commerce product description generation, research on crafting a model so that the vocabulary adjusts to specific consumer segments (i.e. kids) could yield interesting results.

A final remark on the indistinguishability of the GPT-3-generated product descriptions is that contrary to the research of Köbis and Mossink (2021), participants were not financially incentivized to distinguish the GPT-3-generated product descriptions. As a result, participants may have been motivated less and put less effort into distinguishing the GPT-3 generated product descriptions. On the other hand, financially compensating participants makes the research less realistic, as in reality, consumers do not have extrinsic incentives either.

Concerning the examination of CLIP’s performance, this study failed to train a ResNet-50 model on the attribute extraction task. Consequently, the only reference point for the feature extraction capabilities of CLIP was the non-replicated results of the FashionNet model in Liu et al. (2016). Although it was evident that CLIP's performance was inferior to task-specific models, this study was obliged to take results beyond the experimental control for granted instead of keeping the comparison within this study's experimental control.

While this study only explored the effect of finetuning a zero-shot language model in GPT-3, recent publications have shown that finetuning zero-shot image models also offers a promising avenue for increasing the robustness of these models (Wortsman et al., 2021). Since this study showed that CLIP achieves reasonable but inferior performance on apparel category prediction and feature extraction, exploring the effects of finetuning in these areas constitutes encouraging research avenues.

9. References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Chen, Q., Lin, J., Zhang, Y., Yang, H., Zhou, J., & Tang, J. (2019, July). Towards knowledge-based personalized product description generation in e-commerce. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3040-3050).

Das, S. (2020, December 20). How OpenAI's GPT-3 Can Be Alarming For The Society. Analytics India Magazine. https://analyticsindiamag.com/how-openais-gpt-3-can-be-alarming-for-the-society/

Deloitte. (2012). *From bricks to clicks: Generating global growth through eCommerce expansion*. https://www2.deloitte.com/us/en/pages/consumer-business/articles/from-bricks-to-clicks-retail-global-ecommerce.html

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fast, E., & Horvitz, E. (2017, February). Long-term trends in the public perception of artificial intelligence. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2):81.

Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. Journalism, 19(5), 595-610.

Grewal, D., Iyer, G. R., & Levy, M. (2004). Internet retailing: enablers, limiters and market consequences. Journal of business research, 57(7), 703-713.

Gunning, R. (1952). Technique of clear writing.

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., & Gerjets, P. (2021, July). Can Users Distinguish Narrative Texts Written by an Artificial Intelligence Writing Tool from Purely Human Text?. In International Conference on Human-Computer Interaction (pp. 520-527). Springer, Cham.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Huynh, D., & Elhamifar, E. (2020). Fine-grained generalized zero-shot learning via dense attribute-based attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4483-4493).

Irrera, A. (2021, 17 September). More Brits prefer online shopping since pandemic. Reuters. from https://www.reuters.com/world/uk/more-brits-prefer-online-shopping-since-pandemic-2021-09-17/

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019, May). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Kaski, D. (2002). Revision: Is visual perception a requisite for visual imagery?. Perception, 31(6), 717-731.

Khandkar, S. H., & Maurer, F. (2010). A Domain Specific Language to Define Gestures for Multi-touch Applications.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in human behavior, 114, 106553.

Kong, S., Huang, Z., Scott, N., Zhang, Z. A., & Shen, Z. (2019). Web advertisement effectiveness evaluation: Attention and memory. *Journal of Vacation Marketing*, *25*(1), 130-146.

Levin, A. M., Levin, I. P., & Weller, J. A. (2005). A multi-attribute analysis of preferences for online and offline shopping: Differences across products, consumers, and shopping stages. Journal of Electronic Commerce Research, 6(4), 281.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1096-1104).

MacInnis, D. J., & Price, L. L. (1987). The role of imagery in information processing: Review and extensions. Journal of consumer research, 13(4), 473-491.

Maruzani, R. (2021, January 26). *Are You Unwittingly Helping to Train Google's AI Models?* Medium. https://towardsdatascience.com/are-you-unwittingly-helping-to-train-googles-ai-models-f318dea53aee

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. Language Testing, 24(4), 459-488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods, 42(2), 381-392.

McKinney, V., Yoon, K., & Zahedi, F. M. (2002). The measurement of web-customer satisfaction: An expectation and disconfirmation approach. Information systems research, 13(3), 296-315.

McKinsey & Company. (2020, October). *How COVID-19 has pushed companies over the technology tipping point—and transformed business forever*. https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever

Monroe, K. B., & Krishnan, R. (1985). The effect of price on subjective product evaluations. Perceived quality, 1(1), 209-232.

Orendorff, A. (2021). The Ecommerce Fashion Industry in 2021: Statistics, Trends & Strategy. Shopify Plus. https://www.shopify.com/enterprise/ecommerce-fashion-industry

Papers with Code - Papers With Code : Trends. (2021). Paperswithcode. https://paperswithcode.com/trends

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Park, M. (2006). The compensatory effects of pictorial and verbal information for haptic information on consumer responses in non-store shopping environments. The Ohio State University.

Park, E. J., Kim, E. Y., Funches, V. M., & Foxx, W. (2012). Apparel product attributes, web browsing, and e-impulse buying on shopping websites. Journal of Business Research, 65(11), 1583-1589.

Peck, J., Barger, V. A., & Webb, A. (2013). In search of a surrogate for touch: The effect of haptic imagery on perceived ownership. Journal of Consumer Psychology, 23(2), 189-196.

Peck, J., & Childers, T. L. (2003). Individual differences in haptic information processing: The “need for touch” scale. Journal of Consumer Research, 30(3), 430-442.

Peck, J., & Johnson, J. W. (2011). Autotelic need for touch, haptics, and persuasion: The role of involvement. Psychology & Marketing, 28(3), 222-239.

Peck, J., & Shu, S. B. (2009). The effect of mere touch on perceived ownership. Journal of consumer Research, 36(3), 434-447.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta psychologica, 104(1), 1-15.

Qader, R., Jneid, K., Portet, F., & Labbé, C. (2018, November). Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In Proceedings of the 11th International Conference on Natural Language Generation (pp. 254-263).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477-486.

Rodrigues, T., Silva, S. C., & Duarte, P. (2017). The value of textual haptic information in online clothing shopping. Journal of Fashion Marketing and Management.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Sennikova, T. (2021, February 2). Clothes Classification with the DeepFashion Dataset and Fastai | by Tatiana Sennikova | Feb, 2021 | Medium | Towards Data Science. Medium. https://towardsdatascience.com/clothes-classification-with-the-deepfashion-dataset-and-fast-ai-1e174cbf0cdc

Silva, S. C., Rocha, T. V., De Cicco, R., Galhanone, R. F., & Mattos, L. T. M. F. (2021). Need for touch and haptic imagery: An investigation in online fashion shopping. Journal of Retailing and Consumer Services, 59, 102378.

Solaiman, I., & Dennison, C. (2021). Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. *arXiv preprint arXiv:2106.10328*.

Stadtlander, L. M., & Murdoch, L. D. (2000). Frequency of occurrence and rankings for touch-related adjectives. Behavior Research Methods, Instruments, & Computers, 32(4), 579-587.

Statista. (2021, November 1). *Sales of the leading luxury apparel retailers in the U.S. 2019*. https://www.statista.com/statistics/733057/retail-sales-of-luxury-apparel-retailers-in-the-us/

Van der Lee, C., Krahmer, E., & Wubben, S. (2018, November). Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods. In Proceedings of the 11th International Conference on Natural Language Generation (pp. 35-45).

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In Proceedings of the 12th International Conference on Natural Language Generation (pp. 355-368).

Van Miltenburg, E., Clinciu, M. A., Dušek, O., Gkatzia, D., Inglis, S., Leppänen, L., ... & Wen, L. (2021). Underreporting of errors in NLG output, and what to do about it. arXiv preprint arXiv:2108.01182.

Wang, W., Xu, Y., Shen, J., & Zhu, S. C. (2018). Attentive fashion grammar network for fashion landmark detection and clothing category classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4271-4280).

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1-37.

Wortsman, M., Ilharco, G., Li, M., Kim, J. W., Hajishirzi, H., Farhadi, A., ... & Schmidt, L. (2021). Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*.

Zhang, Y., Zhang, P., Yuan, C., & Wang, Z. (2020). Texture and shape biased two-stream networks for clothing classification and attribute recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13538-13547).

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. arXiv preprint arXiv:2102.09690.

Zhou, S., Jeong, H., & Green, P. A. (2017). How Consistent Are the Best-Known Readability Equations in Estimating the Readability of Design Standards? *IEEE Transactions on professional communication, 60(1), 97.*

Appendix

Appendix A

Figure 1: Examples of attributes and bounding boxes

Appendix B

Table 1: Prompt design.

Instruction	Example	Product information	Product Description
One-shot	**Example 1**	product: 'The product is a Zach Straight Leg Jeans. The gender is man. The materials are 69% cotton, 25% elastomultiester, 5% polyester, 1% elastane. The colour is Dark Maui. the jeans has Five-pocket style and Zip fly with button closure	product description: Light whiskers highlight the thighs of comfortable straight-leg jeans cut from soft, dark-washed stretch denim. The jeans has a Five-pocket style and Zip fly with button closure. They're made from 69% cotton, 25% elastomultiester, 5% polyester, and 1% elastane making them perfect for any wearer!
Few-shot	**Example 1**	product: 'The product is a Zach Straight Leg Jeans. The gender is man. The materials are 69% cotton, 25% elastomultiester, 5% polyester, 1% elastane. The colour is Dark Maui. the jeans has Five-pocket style and Zip fly with button closure.	product description: Light whiskers highlight the thighs of comfortable straight-leg jeans cut from soft, dark-washed stretch denim. The jeans has a Five-pocket style and Zip fly with button closure. They're made from 69% cotton, 25% elastomultiester, 5% polyester, and 1% elastane making them perfect for any wearer!
Few-shot	**Example 2**	product: The product is a women's Zip-Up Hoodie. The gender is woman. The materials are 100% organic cotton. The colour is Grey. the hoodie has Hood is lined and Front zip closure"	product description: This indispensable French terry zip hoodie made from soft organic cotton will be your go-to for cozy comfort time and time again. This hoodie is made from cotton and has a hood that's lined. It has a front zip closure to make it more comfortable to wear in your car or on the go!
Few-shot	**Example 3**	product: 'The product is a Garment Dyed Allday Crewneck T-Shirt. The gender is man. The materials are 55% cotton, 35% polyester, 10% viscose. The colour is Dried Cedar. the t-shirt has Crewneck and Short sleeves'	product description: "With sleeves that hit at just the right spot on the biceps, Madewell's signature garment-dyed T-shirt nails the not-too-big, not-too-tight thing. The t-shirt is made from 55% cotton, 35% polyester, and 10% viscose. It has a crewneck design and comes in short sleeves.
Fine-tune	**Example 1**	The product is a Mason Skinny Fit Jeans. The gender is man. The materials are 99% cotton, 1% elastane. The colour is Blue. the jeans has Five-pocket style and Button fly.	Whitewash sanding and whiskering burnish the well-faded blue wash of jeans cut in a modern skinny fit from soft stretch-cotton denim
Fine-tune	**Example 2**	The product is a Dre Low Rise Slim Boyfriend Jeans. The gender is Woman. The materials are 92% cotton, 6% polyester, 2% elastane. The colour is Caliment. the Jeans has Zip fly with button closure and Five-pocket style"	With their perfectly broken-in look and slim fit, these whiskered boyfriend jeans are ready to take on the weekend in the most relaxed way.
Fine-tune	**Example 3**	The product is a Slim Fit Wool Blend Sport Coat. The gender is man. The materials are 73% wool, 25% cotton, 2% elastane. The colour is Black. the Blazer has four-button cuffs and chest pocket	A finely textured wool-blend with a hint of stretch and a slim fit mean versatile appeal for this handsome sport coat.
Fine-tune	**Example 4**	The product is a Mt. Hood Stripe Jacket. The gender is man. The materials are 100% wool. The colour is Black and Olive stripe. the Jacket has Point collar and One-button cuffs	A lightweight wool shell traps warmth inside this striped jacket that's perfect for layering on cool days spent outdoors.
Fine-tune	**Example 5**	The product is a Beck Stretch Performance Shorts. The gender is man. The materials are Polyester/cotton/spandex. The colour is Blue. the Shorts has Zip fly with button closure and Side slant pockets; back welt pockets.	Cheer on those players from the sidelines or jump on in while wearing these smooth subtly-striped shorts designed with four-way-stretch for movement and a quick-dry finish to keep you fresh for dinner after.

Table 2, Male and female product image and prompt design.

Product image	Product image
	
Input prompt	**Input prompt**
The product is a Cigarette Skinny Fit Jeans. The gender is men. The materials are 94% cotton, 4% polyester, 2% elastane. The colour is black. the jeans has Button fly and Five-pocket style.	The product is a Ab-Solution Raw Hem Skinny Jeans. The gender is woman. The materials are 63% cotton, 23% polyester, 11% viscose, 3% Lycra® spandex. The colour is indigo. the jeans has Zip fly with button closure and Five-pocket style.

Table 3, Product description evaluation generated product descriptions

	Male		Female	
	Haptic	**Non-Haptic**	**Haptic**	**Non-Haptic**
One-shot	A smooth black fabric creates slimming straight leg jeans with mid rise and slim, tapered leg. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, and 2% elastane making them perfect for any wearer!	CIGARETTE, the classic skinny fit jean, is developed with a flattering mid-rise waist that's contoured with a waistband that provides a more feminine fit. The jean has a Button fly and Five-pocket style. They're made from 94% Cotton and 4	Feel leather-look in these slim 'n smooth, flattering jeans. These jeans are soft, comfortable and are made from cotton, polyester, viscose, and Lycra® spandex for a sleek, smooth look. These jeans feature zip fly with button closure, five-pocket stylin	These jeans for women provide every element for an amazing fit. The jeans has a zip fly with button closure and five-pocket style. The waist of these jeans is fitted with a belt-looped waistband while the knee are slightly distressed.
Few-shot	These slim-fit, 5-pocket jeans have a narrow leg and a slightly lower rise for a comfortable yet smooth look. This jeans is made from 94% cotton, 4% polyester, and 2% elastane. It has a button fly and a five-pocket style.	Our skinny fit jeans are cut for a close-to-the-body fit without the squeeze, and washed for a worn-in look.The jeans are made from 94% Cotton, 4% Polyester, 2% Elastane. They are Button fly and Five-pocket style.	This tight-fitting pair is designed to slim your tummy and hips. The jeans are made from soft cotton, polyester, viscose, and Lycra® spandex. It has a zip fly with button closure and five-pocket style. They are skinny fit and are available in indigo.	A skinny jean with muscle. Slim, cropped-leg jeans in a stretchy denim with a raw hem. The jeans is made from 63% Cotton, 23% Polyester, 11% Viscose, and 3% Lycra® spandex. It has a zip fly with button closure an
finetune	With their slim fit, soft stretch denim and broken-in feel, these are straight-from-the-moment jeans you'll wear on repeat. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, 2% elastane for a comfortable fit that will keep you looking and feeling great!	A classic for casual days, these jeans have a mid-rise waist and skinny fit featuring artfully ripped knees and a faded black wash. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, 2% elastane fabric.	A slim leg achieves a clean profile in these smooth, dark-washed skinny jeans. The jeans has a Button fly and Five pocket design. It's made from 94 percent cotton, 4 percent polyester, and 2 percent elastane to provide comfort during wear.	A classic for casual days, these black skinny jeans have a roomy, but not baggy, fit and a midrise waistline. The jeans has a Zip fly with button closure and a Five-pocket design. The jeans are made from 63% cotton, 23% polyester, 11% viscose and 3% Lycra® spandex.

Table 4, All Generated Men's product descriptions for human and automatic evaluation.

	Instruction Method			
	Zero shot	**One shot**	**Few shot**	**Finetune**
Sentence 1 with haptic information	Cigarette Skinny Fit Jeans is mens jeans, the height is short. We offer the best service! If you have any questions, please feel free to contact us. the delivery time is 10-15 days. The size is 28-32. Our aim is to make	The Cigarette Skinny Fit jeans by Diesel are inspired by the sharp and fashionable styles prevalent in the 1980s. The jeans is crafted from soft cotton that feels comfortable against the skin and is soft to touch. This slim fit denim is slim from the knee down, and has a straight shape from	This light-wash pair has a clean, classic straight-leg shape that's slim in the hip and roomy in the thigh, with a clean white welt pocket at the front and two at the back. They're made from 94% cotton, 4% polyester, and 2% elastane, making them comfortable to wear for	Stretchy cigarette-style jeans offer a slimming shape in a classic light-wash finish. The jeans has a Button fly and Five pocket design. The fabric is 94% cotton, 4% polyester and 2% elastane for a comfortable fit that will last you through many wears!
Sentence 2 with haptic information	The product is a Cigarette Skinny Fit Jeans. The gender is men. The materials are 94% cotton, 4% polyester, 2% elastane. The colour is black. the jeans has Button fly and Five-pocket styl	A smooth black fabric creates slimming straight leg jeans with mid rise and slim, tapered leg. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, and 2% elastane making them perfect for any wearer!	High-rise skinny jeans are super-flattering, and this pair's soft cotton is comfortable to boot.The jeans are made from 94% cotton, 4% polyester, and 2% elastane. They have a button fly closure and come in a five-pocket styl	A slim leg achieves a clean profile in these smooth, dark-washed skinny jeans. The jeans has a Button fly and Five pocket design. It's made from 94 percent cotton, 4 percent polyester, and 2 percent elastane to provide comfort during wear.
Sentence 3 with haptic information	This product description will be used for e-commerce website. The style of the description is clean and concise. It should be easy to read for the customers. The sentence structure should be simple so that the sentences are short and sweet. There should be different product description for short, tall, ski	Enjoy the wide legline of a straight-leg jean with a slim, straight leg and frayed hem. Cut from comfortable cotton denim with a little stretch and frayed edges, it's a cool-weather essential for your denim line-up. The jeans has Button fly and Fi	These sleek, slim-fitting jeans are polished with a crisp, dark wash. They're made from 94% cotton, 4% polyester, and 2% elastane and have a button fly and five-pocket style.	Light whiskers highlight the legs of comfortable straight-leg jeans cut in a slim fit from soft stretch-cotton denim. The jeans has a Button fly and Five-pocket design. They're made from 94% cotton, 4% polyester, 2% elastane for a comfortable fit that will look great on any woman!
Sentence 4 with haptic information	The product has a slim and skinny fit and a comfortable elastic waist. The material is cotton and elastic and soft and comfortable and sturdy and elastic and comfortable and smooth and comfortable and elastic and solid and sturdy and comfortable and sturdy and elastic and comfortable and comfortable and elastic and elastic and comfortable and comfortable	This slim-fitting style is a cool alternative to straight-leg. Featuring a new Cigarette shape and superskinny cuts, they're perfect for pairing with boots or leather sneakers. They're cut from soft cotton denim with just a hint of stretch for comfort	These slim-fit, 5-pocket jeans have a narrow leg and a slightly lower rise for a comfortable yet smooth look.This jeans is made from 94% cotton, 4% polyester, and 2% elastane. It has a button fly and a five-pocket style. #	With their slim fit, soft stretch denim and broken-in feel, these are straight-from-the-moment jeans you'll wear on repeat. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, 2% elastane for a comfortable fit that will keep you looking and feeling great!
Sentence 5 with haptic information	Cotton jeans. The perfect jeans for guys who want a comfortable and soft pair of jeans with basic features. The cotton fabric is soft and comfortable to wear and has a basic look. Perfect for everyday use and for sport activities such as running. Wear the Cigarette Skinny Fit Jeans w	The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, and 2% elastane making them perfect for any wearer! ### product: The prod	Our slimming, stretchy-yet-sleek, straight-leg jeans are cut to sit just below the waist for a clean look that's never bulky.The jeans are made from 94% cotton, 4% polyester, and 2% elastane. It has a button fly and comes in five-pocket style. It	A slim leg contoured with a curved knee ensures a smooth fit from heel to ankle in sleek, dark-washed jeans that are made for every body type. The jeans has a Button fly and Five-pocket style. The fabric is 94% cotton, 4% polyester, 2% elastane for comfort.
Sentence1 without haptic information	High Quality and Stylish Jeans. The Look is very fashionable and elegant; it is also very charming and attracting. The material is high quality and many people love it. It has a very good quality and is very good jean to wear. the price is acceptable and worth buying	Straight-leg jeans crafted from black stretch denim offer a refined, nod-to-the-'70s look	Made from a lightweight stretch denim, these jeans have a skinny fit through the leg and a zip fly with button closure, five pockets, and a cute, pulled-up hem that makes them a great choice for a night ou	A classic for casual days, these jeans have a mid-rise waist and skinny fit featuring artfully ripped knees and a faded black wash. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, 2% elastane fabric.
Sentence 2 without haptic information	The product is a Cigarette Skinny Fit Jeans. The gender is men. The materials are 94%cotton, 4% polyester, 2% elastane. The colour is bl	Stretch-denim jeans made in the USA of a black twill weave. Features a classic five-pocket design and button fly. Zip fly with button closur	The button-fly, medium-rise Cigarette Jeans from A.P.C. come with a flattering skinny fit. They're crafted in Cone Mills' time-honored White Oak selvage denim—a rarity these day	Pinched-in waist and leg-elongating wide hems lend polish to classic black jeans cut in a modern skinny fit from Italian denim that's a bit stretchy for comfort and durability. The jeans has a Button fly and comes in a variety of sizes. The fabric is 94% cotton, 4% polyester, 2% elastane for a comfortable fit that will look great on any woman who wants to look and feel her best.
Sentence 3 without haptic information	Product description is the most important element of the page. It should be unique, attractive and informative. It should be lengthy enough to include all the necessary details about the product. It should be SEO-friendly. It can consist of sub-sections to include all the details. One sub-sect	CIGARETTE, the classic skinny fit jean, is developed with a flattering mid-rise waist that's contoured with a waistband that provides a more feminine fit. The jean has a Button fly and Five-pocket style. They're made from 94% Cotton an	Our skinny fit jeans are cut for a close-to-the-body fit without the squeeze, and washed for a worn-in look.The jeans are made from 94% Cotton, 4% Polyester, 2% Elastane. They are Button fly and Five-pocket styl	A classic for casual days, these skinny jeans have a roomy, but not baggy, fit and a handsome mid-rise waist. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, 2% elastane for a comfortable fit that will last!
Sentence 4 without haptic information	The product is a Cigarette Skinny Fit Jeans. The gender is men. The materials are 94%cotton, 4%polyester, 2%elastane. The colour is black. The jeans has Button fly and Five-pocket styl	Exuding laid-back attitude, these men's black skinny jeans are crafted in skinny-fit stretch denim and styled in a low-rise hipster silhouette. The jeans have Button fly and Five-pocket style. They're made from 94% Cotton, 4% Polyester, and 2	The jeans are made from 94% Cotton and 4% elastane. They have a Button fly and a Five-pocket style. They're perfect for any occasion and for any weare	A classic for casual days, these are skinny jeans made from stretchy, black denim with a low-rise, ankle-elongating profile. The jeans has a button fly and comes in a five-pocket design. The fabric is 94% cotton, 4% polyester, 2% elastane for comfort during wearability!
Sentence 5 without haptic information	Generate appealing product descriptions for boo	Full stretch, lightweight denim is skinny-fitting through the hip and thigh for a streamlined profile. Features a button fly and five-pocket style. Cut from a blend of 94% Cotton, 4% Polyester and 2% Elastan	These dark-wash jeans are cut with a skinny fit from a stretchy, lightweight denim that's slightly slouchy through the leg.'The jeans is made from 94% Cotton, 4% Polyester, and 2% Elastane. It has a button fly and comes in a five-pocket styl	With their perfectly tapered legs, these are jeans designed for the frame and not the 'fragile' man-boobs. The jeans has a button fly and is made from 94% cotton, 4% polyester, 2% elastane fabric.

Table 5, All Generated Woman's product descriptions for human and automatic evaluation.

Instruction Method

	zero shot	One shot	Few shot	Finetune
Sentence 1 with haptic information	The product is a Ab-Solution Raw Hem Skinny Jeans for WoM. The gender is woman. The materials are 63% cotton, 23% polyester, 11% viscose, 3% Lycra® spandex. The colour is indigo. the jeans has Zi	Feel leather-look in these slim 'n smooth, flattering jeans. These jeans are soft, comfortable and are made from cotton, polyester, viscose, and Lycra® spandex for a sleek, smooth look. These jeans feature zip fly with button closure, five-pocket stylin	Slim your silhouette with these sleek skinny jeans, cut from stretchy cotton-blend fabric and finished with raw hems.The jeans are made from 63% cotton, 23% polyester, 11% viscose, and 3% Lycra spandex. They have a zip fly with button closure and a five-pocket styl	Stretchy soft denim flexes and retains shape on classic skinny jeans with a roomy, lived-in feel and a clean finish that's easy to layer. The jeans has a Zip fly with button closure and a Five-pocket design. They're made from 63% cotton, 23% polyester, 11% Viscose, 3% Lycra® spandex.
Sentence 2 with haptic information	Made from soft cotton fabric, these iconic jeans feature a slim fit and clean design for a timeless design that goes with anything in your wardrobe. Finished with rolled hem for an edgy touch'	Cotton/polyester/spandex blend. Imported.	These jeans are the perfect way to round out any wardrobe. With a flattering straight fit that's just tight enough at the ankle, they're a wardrobe essential.The jeans have a zip fly with a button closure, five pockets, and a straight fit. They're made from 63% cotton, 23% polyester, 11% viscose	A slim leg achieves a clean profile in these raw-washed skinny jeans designed with a "mom" wash finish and a high rise waist. The jeans has a Zip fly with a button closure and has Five pockets for convenience. The material is 63% cotton, 23% polyester, 11% Viscose and 3% Lycra® spand
Sentence 3 with haptic information	The descriptions tell you that the product is suitable for working out, partying, dating, clubbing, shopping, traveling, daily wear. The product is made in USA. The measurements are in inches. The product is slim fit. The measurements are in inches'	The jeans have zip fly with button closure. They're made from 63% cotton, 23% polyester, 11% viscose and 3% Lycra® spandex. They're comfortable, flattering, and versatile.	This tight-fitting pair is designed to slim your tummy and hips. The jeans are made from soft cotton, polyester, viscose, and Lycra® spandex. It has a zip fly with button closure and five-pocket style. They are skinny fit and are available in indigo.	A light whisker outlines the ankle bones of delicate high-waist jeans cut in a smooth, flat-felled fit and finished with a raw hemline. The jeans has a Zip fly with a button closure and is made from 63% cotton, 23% polyester, 11% of a material called "viscose." The jeans will fit most people well as they have a waist of 34 inches and a leg of 36 inches.
Sentence 4 with haptic information	Fashionable jeans that shape and smooth the body. The cotton fabric and Lycra® spandex fabric gives soft and comfortable touch and flexible fit. Five-pocket style jeans with button closure and zip fly'	Ab-Solution Raw Hem Skinny Jeans from River Island. The jeans has a Zip fly with button closure and Five-pocket style. They're made from 63% cotton, 23% polyester, 11% viscose, and 3% Lycra® spandex.	The Ab-Solution Raw Hem Skinny Jeans have a zip fly with button closure and five-pocket style. They're made from 63% cotton, 23% polyester, 11% viscose, and 3% Lycra® spandex.	Ab-solution technology makes you feel instantly dry and comfortable in these smooth, stretch denim jeans designed for the shape of your body. The jeans has a Zip fly with button closure and a Five pocket design. The jeans are made from 63% cotton, 23% polyester, 11% Viscose and 3% Lycra® spandex fabric to provide comfort during wearability!
Sentence 5 With haptic information	Stylish skinny jean with worn look.Exposed button closure. Zip fly with button closure. Five-pocket style. High-rise. Straight-leg. Cotton-polyester-spandex. Machine wash cold, line dry. Made in the USA of imported fabric. Measurement	These slim-fitting jeans are cut in soft, mediumweight cotton and detailed with fading, whiskering and intricate stitching. Style Name:Ab-Solution Raw Hem Skinny Jeans. Style Number: 5356320. Available in stores.	The slim, straight cut of these raw-hem skinny jeans is streamlined and sharp, with a cool hand-destroyed look thanks to a little bleach action.The jeans are made from 63% cotton, 23% polyester, 11% viscose, and 3% Lycra spandex. They have a zip fly with button closur	A slim leg achieves a clean profile in these light-wash skinny jeans designed with a high rise waist and functional back pocket for your phone. The jeans has a Zip fly with a button closure and a Five-pocket design. The fabric is 63% cotton, 23% polyester, 11% Viscose and 3% Lycra® spandex to provide comfort while wearing!
Sentence 1 without haptic information	The Ab-Solution Raw Hem Skinny Jeans features a high-rise waist that measures at the belly button. The jeans feature an ankle length. The style of the jeans is a skinny jean. The hem of the jeans is a raw hem. The jeans feature a zip fly with butto	Slim jeans cut from a supersoft lightweight denim are styled with a raw hem, frayed edges and welt front pockets. They're made from a stretchy, modern-feeling denim that moves with you, and feature a zip fly with button closure and a five-pocket style.	Treat yourself to a pair of skinny jeans that won't fade fast. The Ab-Solution Raw Hem Skinny Jeans are made from 63% organic Cotton, 23% polyester, 11% viscose and 3% Lycra® spandex. They have a zip fly with button closure, a five-pocket style, an	A bright blue wash furthers the faded-wash look of jeans shaped from the waist to the knees in a distressing raw-hemp finish. The jeans has a Zip fly with a button closure and a Five-pocket design. The denim is made from 63% cotton, 23% polyester, 11% of the material is a material called "viscose." The jeans are available in four different colors!
Sentence 2 without haptic information	The product is a Ab-Solution Raw Hem Skinny Jeans. The gender is woman. The materials are 63% Cotton, 23% Polyester, 11% Viscose, 3% Lycra® Spandex. The colour is indigo. The jeans has Zip fly wit	The Ab-Solution Raw Hem Skinny Jeans by True Religion are your new go-to jeans. They're made from 63% Cotton, 23% Polyester, 11% Viscose, and 3% Lycra® Spandex, and feature Zip fly with button closure, an	With a high-rise waist that skims the body and a skinny cut that hugs the hips and thighs, these versatile, medium-wash jeans are a must-have in any wardrobe.The jeans are made from 63% Cotton, 23% Polyester, 11% Viscose, and 3% Lycra® spandex	A classic for casual days, these low-rise jeans feature a premium stretch denim construction and a classic raw hem. The jeans has a Zip fly with button closure and is made from 63% cotton, 23% polyester, 11% of a material called "viscose." The jeans have five pockets, giving you plenty of room for all of your needs!
Sentence 3 without haptic information	The product is a Ab-Solution Raw Hem Skinny Jeans. The gender is woman. The materials are 63% cotto, 23% polyester, 11% viscose, 3% Lycra® spandex. The colour is indigo. The jeans has Zip fly wit	These jeans are made from 63%cotton,23%polyester,11%viscose,3%Lycra®spandex.They are perfect for any wearer.Light whiskers highlight the thighs of these jeans designed to a skinny fit from stretch denim. The jeans have	These jeans have a zip fly with button closure, five-pocket style, and an Ab-Solution Raw Hem Skinny fit. The jeans are made from 63% Cotton, 23% polyester, 11% viscose, and 3% Lycra® spandex.	A classic for casual days, these black skinny jeans have a roomy, but not baggy, fit and a midrise waistline. The jeans has a Zip fly with button closure and a Five-pocket design. The jeans are made from 63% cotton, 23% polyester, 11% of a material called "viscose," and 3% Lycra® spandex.
Sentence 4 without haptic information	The product is a Ab-Solution Raw Hem Skinny Jeans. It's a woman's jeans. The materials are 63% Cotton. 24% Polyester. 11% Viscose. 4% Lycra® spandex. The colour is indigo. There's a Zi	These jeans for women provide every element for an amazing fit. The jeans has a zip fly with button closure and five-pocket style. The waist of these jeans is fitted with a belt-looped waistband while the knee are slightly distressed.	A skinny jean with muscle.Slim, cropped-leg jeans in a stretchy denim with a raw hem. The jeans is made from 63% Cotton, 23% Polyester, 11% Viscose, and 3% Lycra® spandex. It has a zip fly with button closure an	Made from a super-lightweight stretch denim, these flattering jeans have a ringer neck and raw hem for plenty of lived-in character. The jeans has a Zip fly with button closure and is made from 63% cotton, 23% polyester, 11% of the material is a material called "viscose." The jeans are perfect for any woman who wants her jeans to look and feel their best!
Sentence 5 without haptic information	With a super-slim fit from hip to hem, this modern take on a classic 5-pocket jeans boasts a cropped rise for a look that will make the most of your most enviable assets. In a classic, dark wash,'	These jeans have a flattering high-rise that's worn with a raw-hemmed, frayed hem for a deconstructed touch. Zip fly with button closure. Five-pocket style.	Skinny jeans with a raw hem give your look a grunge-meets-glam edge. The jeans is made from 63% Cotton, 23% polyester, 11% Viscose, and 3% Lycra® spandex. It has a zip fly with button closure and Five-pocket style.	A modern fit and easy-to-absorb stretch denim bring you closer to nature in these skin-tight jeans. The jeans has Zip fly with button closure and Five-pocket style. The jeans is made from 63% cotton, 23% polyester, 11% of a material called "viscose." This material has a lot of give, but also holds its shape well so the jeans will look great o

Table 6, Readability and Lexical Diversity Scores Men's product descriptions.

Method	Flesch	Gunning	MTLD	HD-D
Zero shot	65.915	8.966	30.8318	0.3781
	(49.4382, 82.3918)	(4.8399, 13.0921)	(19.1088, 42.5537)	(0.1318, 0.6243)
One shot	76.581	6.602	70.2432	0.0727
	(71.5308, 81.6312)	(5.021, 8.183)	(38.5164, 101.97)	(-0.0917, 0.2371)
Few shot	75.679	8.862	58.5876	0.2541
	(68.8503, 82.5077)	(6.7968, 10.9272)	(39.4339, 77.7414)	(-0.0388, 0.547)
Finetune	78.139	7.721	84.4213	0.6848
	(74.0853, 82.1927)	(6.52, 8.922)	(70.6586, 98.1841)	(0.4258, 0.9432)

Note. 95% Confidence intervals in parenthesis.

Table 7, Readability and Lexical Diversity Scores Women's product descriptions.

Method	Flesch	Gunning	MTLD	HD-D
Zero shot	74.729	6.677	48.1518	0.0546
	(69.4359, 80.0221)	(5.8647, 7.4893)	(33.7066, 62.5969)	(-0.069, 0.1782)
One shot	62,496	8.024	76.2566	0.0748
	(40.3564, 84.6356)	(6.087, 9.961)	(33.0308, 119.4824)	(-0.0943, 0.2438)
Few shot	74.637	8.635	97.4594	0.5905
	(69.2724, 80.0016	(6.0987, 11.1713)	(60.81, 134.1087)	(0.2983, 0.8826)
Finetune	75.076	8.319	80.6193	0.8279
	(70.4099, 79.7421)	(7.2823, 9.3557)	(69.1151, 92.1235)	(0.795, 0.8608)

Note. 95% Confidence intervals in parenthesis.

Table 8, examples of identified errors in product descriptions.

Error	Model	Product hallucination error
Product hallucination	Zero-shot	Stylish skinny jean with worn look. Exposed button closure. Zip fly with button closure. Five-pocket style. High-rise. Straight-leg. Cotton-polyester-spandex. Machine wash cold, line dry. Made in the USA of imported fabric. Measurement
Product hallucination	One-shot	These jeans for women provide every element for an amazing fit. The jeans has a zip fly with button closure and five-pocket style. The waist of these jeans is fitted with a belt-looped waistband while the knee are slightly distressed.
Product hallucination	Few-shot	This light-wash pair has a clean, classic straight-leg shape that's slim in the hip and roomy in the thigh, with a clean white welt pocket at the front and two at the back. They're made from 94% cotton, 4% polyester, and 2% elastane, making them comfortable to wear for
Product hallucination	Finetune	Pinched-in waist and leg-elongating wide hems lend polish to classic black jeans cut in a modern skinny fit from Italian denim that's a bit stretchy for comfort and durability. The jeans has a Button fly and comes in a variety of sizes. The fabric is 94% cotton, 4% polyester, 2% elastane for a comfortable fit that will look great on any woman who wants to look and feel her best.
Factual incorrectness	One-shot	A smooth black fabric creates slimming straight leg jeans with mid-rise and slim, tapered leg. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, and 2% elastane making them perfect for any wearer!
Factual incorrectness	Few-shot	The jeans are made from 94% Cotton and 4% elastane. They have a Button fly and a Five-pocket style. They're perfect for any occasion and for any wearer!
Factual incorrectness	Finetune	A classic for casual days, these black skinny jeans have a roomy, but not baggy, fit and a midrise waistline. The jeans has a Zip fly with button closure and a Five-pocket design. The jeans are made from 63% cotton, 23% polyester, 11% of a material called "viscose," and 3% Lycra® spand
Copy of data Input	Zero-shot	The product is a Ab-Solution Raw Hem Skinny Jeans. The gender is woman. The materials are 63% cotto, 23% polyester, 11% viscose, 3% Lycra® spandex. The colour is indigo. The jeans has Zip fly wit
Task hallucination	Zero-shot	Product description is the most important element of the page. It should be unique, attractive and informative. It should be lengthy enough to include all the necessary details about the product. It should be SEO-friendly. It can consist of sub-sections to include all the details. One sub-sect
Repetition	Zero-shot	Cotton jeans. The perfect jeans for guys who want a comfortable and soft pair of jeans with basic features. The cotton fabric is soft and comfortable to wear and has a basic look. Perfect for everyday use and for sport activities such as running. Wear the Cigarette Skinny Fit Jeans wit
Unfinished sentence	Zero-shot	Fashionable jeans that shape and smooth the body. The cotton fabric and Lycra® spandex fabric gives soft and comfortable touch and flexible fit. Five-pocket style jeans with button closure and zip fly'
Unfinished sentence	One-shot	These jeans are made from 63%cotton,23%polyester,11%viscose,3%Lycra®spandex.They are perfect for any wearer. Light whiskers highlight the thighs of these jeans designed to a skinny fit from stretch denim. The jeans have
Unfinished sentence	Few-shot	The slim, straight cut of these raw-hem skinny jeans is streamlined and sharp, with a cool hand-destroyed look thanks to a little bleach action. The jeans are made from 63% cotton, 23% polyester, 11% viscose, and 3% Lycra spandex. They have a zip fly with button closur
Unfinished sentence	Finetune	With their perfectly high rise and ultra-tight fit, these are practically blank-slate skinny jeans. The jeans has Zip fly with button closure and Five-pocket style. They're made from 63% cotton, 23% polyester, 11% of a material called "viscose," and 3% of Lycra® spand

Table 9, readability, correctness, and repetition scores split on gender and haptic information

	Readability				Correctness				Repetition			
	Men		Woman		Men		Woman		Men		Woman	
Model	Haptic	Non-Haptic	Haptic	Non-Haptic	Haptic	Non-haptic	Haptic	Non- haptic	Haptic	Non- Haptic	Haptic	Non-haptic
One-shot	4.0588	2.9411	3.6	4.8333	5	3.7647	4.5	5.5	2.8235	2.5294	2.3333	3
Few-shot	5.0588	4.9411	5.125	5.1667	5.2941	5.6471	5.8333	5.1667	3.3529	2.6470	4.3333	3
Fine-tune	5.7647	5.2353	5.626	5.1667	4.0588	4.0588	5.8333	5.1667	2.2941	2.1765	2.3333	2.1667

Table 10, Readability, Correctness, and Repetition scores split on gender

	Readability		Correctness		Repetitiveness	
	Men	Woman	Men	Woman	Men	Woman
One-shot	3.5	4.9166	4.3823	5.	2.6765	2.6667
Few-shot	5	4.75	5.4706	5.5	3	3.6667
Fine-tune	5.5	5.3333	4.2941	5.5	2.2353	2.25

Appendix C – Product Description Evaluation

Male Product Image - one shot description [Low NFT]

CIGARETTE, the classic skinny fit jean, is developed with a flattering mid-rise waist that's contoured with a waistband that provides a more feminine fit. The jean has a Button fly and Five-pocket style. They're made from 94% Cotton and 4

The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Male Product Image - few shot description [Low NFT]

Our skinny fit jeans are cut for a close-to-the-body fit without the squeeze, and washed for a worn-in look.The jeans are made from 94% Cotton, 4% Polyester, 2% Elastane. They are Button fly and Five-pocket style.

The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the product materials feel like.

This product description contains words that describe what touching the product would feel like.

Male Product Image - fine-tune description [Low NFT]

A classic for casual days, these jeans have a mid-rise waist and skinny fit featuring artfully ripped knees and a faded black wash. The jeans has a Button fly and Fivepocket style. They're made from 94% cotton, 4% polyester, 2% elastane fabric. The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Male product image - few shot description [High NFT]

These slim-fit, 5-pocket jeans have a narrow leg and a slightly lower rise for a comfortable yet smooth look. This jeans is made from 94% cotton, 4% polyester, and 2% elastane. It has a button fly and a five-pocket style.

The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the product materials feel like.

This product description contains words that describe what touching the product would feel like.

Male product image - fine-tune description [High NFT]

With their slim fit, soft stretch denim and broken-in feel, these are straight-from-themoment jeans you'll wear on repeat. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, 2% elastane for a comfortable fit that will keep you looking and feeling great!

The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Male Product Image - One shot description [High NFT]

A smooth black fabric creates slimming straight leg jeans with mid rise and slim, tapered leg. The jeans has a Button fly and Five-pocket style. They're made from 94% cotton, 4% polyester, and 2% elastane making them perfect for any wearer! The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Female product image - one-shot description [Low NFT]

These jeans for women provide every element for an amazing fit. The jeans has a zip fly with button closure and five-pocket style. The waist of these jeans is fitted with a belt-looped waistband while the knee are slightly distressed.

The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Female product image - few-shot description [Low NFT]

A skinny jean with muscle. Slim, cropped-leg jeans in a stretchy denim with a raw hem.
The jeans is made from 63% Cotton, 23% Polyester, 11% Viscose, and 3%
Lycra® spandex. It has a zip fly with button closure an
The text in this product description flows in a natural manner and is easy to read.

Based on the product image, the product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the product materials feel like.

This product description contains words that describe what touching the product would feel like.

Female product image - fine-tune description [low NFT]

A classic for casual days, these black skinny jeans have a roomy, but not baggy, fit and a midrise waistline. The jeans has a Zip fly with button closure and a Five-pocket design. The jeans are made from 63% cotton, 23% polyester, 11% viscose and 3% Lycra® spandex.

The text in this product description flows in a natural manner and is easy to read.

The product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Female product image - fine-tune description [High NFT]

A slim leg achieves a clean profile in these smooth, dark-washed skinny jeans. The jeans has a Button fly and Five pocket design. It's made from 94 percent cotton, 4 percent polyester, and 2 percent elastane to provide comfort during wear.

The text in this product description flows in a natural manner and is easy to read.

The product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Female product image - few shot description [High NFT]

This tight-fitting pair is designed to slim your tummy and hips. The jeans are made from soft cotton, polyester, viscose, and Lycra® spandex. It has a zip fly with button closure and five-pocket style. They are skinny fit and are available in indigo.

The text in this product description flows in a natural manner and is easy to read.

The product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Female product Image - one shot description [High NFT]

Feel leather-look in these slim 'n smooth, flattering jeans. These jeans are soft, comfortable and are made from cotton, polyester, viscose, and Lycra® spandex for a sleek, smooth look. These jeans feature zip fly with button closure, five-pocket stylin The text in this product description flows in a natural manner and is easy to read.

The product description includes correct information.

The product description contains unnecessary repeated information.

This product description contains words that describe how the products material feels like.

This product description contains words that describe what touching the product would feel like.

Appendix D – Purchase Intention Experiment

Woman - High NFT - AI - Blue Jeans

A slim leg achieves a sleek profile in these cropped straight jeans designed with a high rise waist. This jeans has a zip fly with a button closure and a Five pocket design. The jeans are made from 99% cotton and 1% elastane for comfort during wearability!

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Woman - High NFT - AI - Black Hoodie

A slouchy fit brings soft structure to a versatile cotton-blend hoodie with a bit of stretch for shape and comfort. The hoodie has a Drawstring hood and long sleeves with ribbed cuffs. It's made from 64% cotton, 36% polyester for comfort wearability in a variety of situations!

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Woman - Low NFT - Human - Grey Hoodie

A slouchy oversized fit lends a borrowed-from-him look to a cozy pullover hoodie made from an eco-friendly recycled-cotton blend. The hoodie features a kangaroo pocket and it is made from 80% organic cotton, and 20% polyester.

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Woman - Low NFT - Human - Blue Jeans

Rock a modern casual-cool silhouette in these high-waist jeans punctuated with elongating wide legs. They are made from 99% cotton and 1% spandex, and the medium dark indigo wash creates a wear-with-everything versatility!

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- ○ Written by a human
- ○ Generated by AI

Woman - High NFT - Human - Black Hoodie

Layer up with comfort at the gym, at home or while running errands in a classic cotton-blend hoodie styled with logoembroidered lettering at the chest. The hoodie is made from a 65% cotton and 35% polyester blend, which keep it looking good and feeling soft for a long time.

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Woman - High NFT - Human - Blue skinny jeans

New Roadtripper Authentic denim brings an old-school, rigid look and tons of stretch to these comfy skinny jeans reengineered for those with an hourglass shape. Made from a blend of 76% cotton, 22% recycled polyester, and 2% spandex, these jeans provide a comfortable yet stylish fit.

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Woman - Low NFT - AI - Grey hoodie

A slouchy kangaroo pocket covers this cozy, drawstring hoodie that's ideal for relaxing at home or heading to the gym. The hoodie has drawstring hood and comes in a pullover style. The fabric is 50% polyester, 46% cotton, 4% rayon for comfort in any weather!

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- ○ Written by a human
- ○ Generated by AI

Woman - Low NFT - AI - light blue Jeans

Soft light-blue wash furthers the casual, comfortable feel of these slim-tailored jeans cut from Italian denim that's so comfortable you'll never want to take them off. This jeans has a Zip fly with button closure and it comes in a Five pocket style. It's made from 52% cotton, 34% lyocell, 12% polyester, and 2% spandex.

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Man - High NFT - AI - Black skinny jeans

A slim leg achieves a clean profile in these smooth, dark-washed skinny jeans that are made for every body type. The jeans has a button fly and comes in a five pocket style. It's made with 94% cotton, 4% polyester and 2% elastane for a comfortable fit that will last!

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Man - High NFT - AI Hoodie Black

A soft cotton-blend drawstring hoodie with a smooth finish and curved hemline is a comfortable way to warm up on a chilly day. The hoodie has a Drawstring hood and ribbed cuffs and hem. It's made from 80% cotton and 20%
polyester for comfort wearability in any season!

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Man - Low NFT - AI - blue jeans

A classic for casual days, these faded blue jeans have a mid-rise waist and are designed with a bit of stretch for versatility. The jeans has a Zip fly with button closure and Five pocket style. They're made from 98% elastane and 2% cotton to provide comfort during wearability.

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- ○ Written by a human
- ○ Generated by AI

Man - Low NFT - AI - Hoodie Grey

With drop sleeves and loose-fitting profile, this fleece hoodie will keep you cozy and dry in stormy conditions or while you're hanging with friends. The hoodie has a drawstring hood and kangaroo pocket. It's made from 100% cotton for a comfortable fit!

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Man - Low NFT - Human - Grey Hoodie

Crafted from a organic-cotton blend, this hoodie is ready for the weekend. The hoodie has drawstring hood and a kangaroo pocket, making it a wardrobe essential! It is made from 80% organic cotton, and 20% polyester.

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Man - Low NFT - Human - blue skinny jeans

Sanded and faded to old-favorite perfection, these contour-hugging jeans sport plenty of stretch and the enhanced comfort of sustainably produced cotton. This five-pocket style jeans has a zip fly with button closure, and is made from 95% cotton, 4% elastomultiester, and 1% elastane which ensures shape preservation.

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- ○ Written by a human
- ○ Generated by AI

Man - High NFT - Human - grey jeans

Sanding at the puckers and wear points burnishes the faded wash of jeans cut slim with a modern straight-leg profile that looks great day or night. The jeans have a zip fly with button closure, and they are made from 51% rayon, 26% cotton, 22% polyester, and 1% spandex, to provide you with everyday comfort!

If you needed a new pair of jeans, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI

Man - High NFT - Human - Black Hoodie

This easy hoodie knit from a soft cotton blend is an essential layering piece for the daily rotation. It features a kangaroo pocket and is made from 75% cotton and 25% polyester.

If you needed a new hoodie, would you order this product to try it out?

Please describe your overall feeling about the product

This product description was...

- Written by a human
- Generated by AI