Appendix for "Synthetic Product Review Generation for Market Analysis: A GPT-2 and LLaMA 3 Approach"

A Additional Information

A.1 Technical Information on GPT-2 and LLaMA 3

OpenAI's GPT-2 Medium (Radford et al., 2019) is a Transformer-based (Vaswani et al., 2017) decoder-only (unidirectional) generative pre-trained LM (PLM) with 355M parameters that was pre-trained on 40GB of data. Other features of the PLM include the byte-pair encoding (BPE) tokenization algorithm (Sennrich et al., 2016) and pre-normalization (layer normalization of the input layer).[1]

Meta AI's LLaMA 3 8B (Dubey et al., 2024) is also a Transformer-based (Vaswani et al., 2017) decoder-only (unidirectional) generative PLM, but it differs from GPT-2 with regard to several key aspects. First, it is a much larger model with 8B parameters. Second, it is multilingual, supporting eight languages. Third, it is multimodal, meaning that it can generate not only text but also images, audio, and video. Fourth, it was pre- *and* post-trained with a vast amount of high quality data ($\approx$ 55TB[2] of data).[3]

Both GPT-2 and LLaMA 3 are based on the Transformer architecture introduced by Vaswani et al. (2017) and illustrated in Figure I.

The self-attention mechanism, the most crucial component of the Transformer architecture, can be expressed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q is the query matrix of an input text, K is the key matrix of an input text, V is the value matrix of an input text, and d_k is the dimensionality of the key vector (Vaswani et al., 2017).

A.2 Prompting

A.2.1 Zero-Shot vs Few-Shot Prompting

Zero-shot prompting, introduced in (Radford et al., 2019), is a prompting method, where the prompt directly instructs a model to perform a task without providing any examples or demonstrations, meaning the model must leverage its pre-existing knowledge to generate an answer (Sahoo et al., 2024).

Few-shot prompting, introduced in (Brown et al., 2020), is a prompting method, where the prompt contains an instruction *and* a few examples to induce a better understanding of the task and generate a better response (Brown et al., 2020; Sahoo et al., 2024). It is called "few-shot" because the model is given a small number of examples ("shots"). Since zero-shot prompting can, in certain scenarios, outperform few-shot prompting (Reynolds & McDonell, 2021), we also use zero-shot prompting.

[1] Additional technical details about the features of GPT can be found in OpenAI's first GPT paper (Radford et al., 2018).

[2] The LLaMA 3 paper states that 15T tokens were used. Assuming one token is 4 bytes, the total size of the training data is 15T tokens $\times$ 4 bytes $\div$ 1024^3 bytes/GB $\approx$ 55TB.

[3] Additional technical details about the features of LLaMA 3 are available in the first two LLaMA papers (Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023).

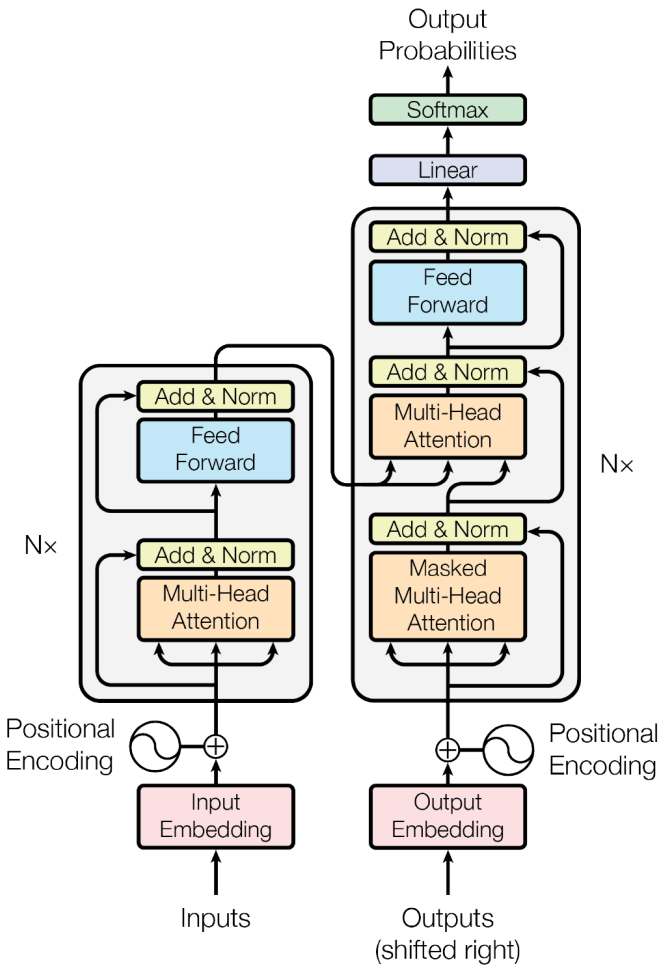

Figure I: Transformer architecture (Vaswani et al., 2017)

Finally, we did not use more advanced prompting techniques, such as chain-of-thought (CoT; Wei et al., 2022), self-consistency (Wang et al., 2023), ReAct (Yao et al., 2023), and retrieval augmented generation (RAG; Lewis et al., 2020), because our task does not require reasoning or (complex) problem-solving and because GPT-2 Medium is less capable than other SOTA LMs. A more advanced prompting strategy could be used for LLaMA 3, but this could lead to large differences in the quality of the synthetic product reviews between GPT-2 and LLaMA 3.

A.2.2 Prompt Engineering

GPT-2 Prompt Generation

Our prompt engineering approach for GPT-2 comprised experimenting with different types of zero-shot prompts in order to obtain coherent and relevant reviews. As a result of fine-tuning, our GPT-2 model could generate review-like content from even a simple prompt, such as "T-shirt."

To improve the quality of our synthetically generated reviews, we began by providing the start of a sentence, as the model is designed to predict the next words based on the preceding text, a method known as causal language modeling (CLM; Bengio et al., 2003; Vaswani et al., 2017).

We first used the prompt "The T-shirt is" and noticed that the first predicted word was often the same across different outputs. Another issue we encountered was that all the reviews began in a similar manner,

which detracts from the realism typically seen in actual reviews and reduces overall variation. This poses a challenge for our goal of conducting market analysis using realistically generated reviews. To address this, we experimented with an instruction-based approach, using a prompt such as "Generate a product review about a T-shirt. Review:", where the "Review:" at the end was intended to serve as a starting point for the model's predictions. While this method resulted in longer and more varied reviews, the content tended to be unfocused. Although the topics were related to product reviews, the lack of logical structure made the reviews seem unrealistic. It appears that since the model is already capable of generating a review structure on its own, adding explicit instructions like "Generate a product review" may be unnecessary and could contribute to this lack of coherence.

Thus, we decided to continue using the statement-based approach, using simple prompts like "I bought a T-shirt." This method produced clearer, more concise reviews. Although they were not perfect yet, the quality is noticeably better compared to reviews generated by other prompts.

Additionally, the prompting literature suggests that providing context to the model can improve the quality of generated responses (Bsharat et al., 2023; Lo, 2023). Therefore, in our next step, we aimed to improve our results by incorporating additional contextual information into the prompt. It is crucial, however, to avoid introducing implicit bias, ensuring the model generates neutral and unbiased responses. The prompt should not suggest specific aspects of the T-shirt, such as quality or delivery, as this could skew the results. This is especially important in market analysis, where the goal is to obtain objective and reliable insights. For this reason, a shorter prompt might be more suitable for generating product reviews in our case. An example of a context-based prompt we used was: "You can find a wide variety of T-shirts on Amazon. I recently ordered one, and it has finally arrived, allowing me to see it in person." However, when generating product reviews, the added context did not seem to improve quality compared to simpler prompts. This could be due to the fine-tuning process, which has already optimized the model for review generation, making the additional context unnecessary.

We also explored an alternative approach, focusing less on the format of the text, such as a review, and more on the content we aimed to generate. This led us to consider what individuals typically express in their reviews, often reflecting on a T-shirt they have purchased and worn a few times. Drawing on previous prompt experiments, we kept the prompt concise: "I purchased this T-shirt on Amazon recently. Here's my experience so far." The period at the end was crucial for guiding the model to start with a complete sentence rather than continuing an incomplete one. This approach proved effective, particularly because it provided a clear starting point for the model. The reviews generated with this prompt were of appropriate length and contained relevant insights about the T-shirts.

Although not all reviews remained strictly focused on T-shirts – occasionally shifting to related garments like dresses or jackets – this may be due to the model's reliance on its training data, which sometimes makes it difficult to clearly differentiate between similar items. However, this prompt significantly reduced such occurrences and consistently produced the highest-quality reviews. These findings align with the perplexity scores calculated for 100 reviews generated by each prompt, as shown in Table I (for the formula of perplexity, see Section A.4). Consequently, we chose to use the prompt *"I purchased this T-shirt on Amazon recently. Here's my experience so far."* to generate our sample of 10,000 reviews.

It is important to note that the prompts discussed above are examples used to illustrate our approach in identifying a suitable prompt. We also experimented with variations of these prompts across different approaches to better understand the model and determine the most effective strategy. The full set of prompts, along with 100 generated reviews and their corresponding perplexity scores, are provided in GitHub.

Table I: Selected Prompts and their Average Perplexities for GPT-2

Prompt	**Perplexity**
"T-shirt"	30.993
"The T-shirt is"	15.950
"Generate a product review about a T-shirt. Review:"	18.276
"I bought a T-shirt."	13.065
"You can find a wide variety of T-shirts on Amazon. I recently ordered one, and it has finally arrived, allowing me to see it in person."	21.668
"I purchased this T-shirt on Amazon recently. Here's my experience so far."	12.972

Note: Perplexity refers to the average perplexity across a sample of 100 synthetic product reviews generated with the given prompt.

LLaMA 3 Prompt Generation

Our prompt engineering approach for LLaMA 3 comprised experimenting with different types of few-shot prompts in order to obtain coherent and relevant reviews. One example of a prompt we tried is:

"The following is a customer review for a T-shirt:

I love this T-shirt! The fabric is incredibly soft, and it fits just right. Even after several washes, it looks as good as new. Highly recommend!

The following is another customer review for a T-shirt: "

We employed one-shot prompting, a variant of few-shot prompting that uses $k = 1$ example(s), as incorporating additional examples would significantly increase the prompt length and introduce computational challenges. One-shot prompting performs comparably to few-shot prompting, depending on the task (Brown et al., 2020). To mitigate potential bias(es) introduced by the prompt, we generated half of the product reviews using a negative example and the other half using a positive example.

Since the model is not fine-tuned, generating sufficiently long individual reviews to extract useful information presents a challenge. Unlike our fine-tuned GPT-2 model, LLaMA 3 is a generalized model, requiring longer prompts (i.e., prompts containing examples of comparable length to the desired output) to produce clear and concise reviews. Initial attempts to directly instruct the model to generate reviews were ineffective, as the resulting outputs often contained hallucinations, such as repeated information and non sequiturs.

Few-shot prompting has inherent limitations. The generated reviews often exhibit significant similarity, repeatedly relying on a few key phrases. Additionally, the outputs show a strong positivity bias (cf. Boucher & Osgood, 1969; Dodds et al., 2015), affecting both the negative and positive example prompts. Furthermore, the model frequently produces multiple reviews, an issue we were unable to resolve despite instructions to generate exactly one review.

Considering our goal was to get the model to generate exactly one review to compare to both the real reviews and the synthetic reviews generated by our GPT-2 model, we achieved this by introducing an asterisk to mark the beginning and end of each review in the prompt, making it possible to extract only the first of the generated reviews. In our statistical evaluation and market analysis, we used these shortened outputs, while we used the entire output for computing perplexity.

We identified the optimal prompts by evaluating the average perplexity of each prompt (for the formula of perplexity, see Section A.4), experimenting with various instructions and example reviews. The best-performing example reviews resulted in the following prompts, which we then used for the generation of synthetic product reviews:

Positive: *"Here is a review of a T-shirt: *This T-shirt is fantastic! The fabric is high-quality, and it's so comfortable to wear all day. I love that it's both lightweight and durable, and the color stays vibrant even after several washes. I'm very impressed and will definitely be buying more!* Another review of a T-shirt:*"*

Negative: *"Here is a review of a T-shirt: *I had high hopes for this T-shirt, but it fell short in every way. The material feels low-quality and rough, and the seams are already coming undone. It also shrank significantly after the first wash, making it unwearable. Save your money and avoid this one.* Another review of a T-shirt: *"*[4]

We tested 10 prompts, generating 5 reviews for each, and calculated the mean perplexity of the outputs. The prompts are documented in Table II and in the corresponding CSV file in GitHub (see here). Additionally, we experimented with varying instructions preceding the examples.

Table II: Selected Prompts and their Average Perplexities for LLaMA 3

Prompt	**Perplexity**
"Here is a review of a T-shirt: *I absolutely love this T-shirt! The material is so soft, and it feels amazing against the skin. I've washed it multiple times, and it still looks brand new. The fit is perfect, and I get compliments every time I wear it. This has become my favorite go-to shirt for both casual and semi-casual outings. Highly recommended!* Another review of a T-shirt:*"	3.627
"Here is a review of a T-shirt: *I can't get enough of this T-shirt! The fabric is so soft and comfortable that it feels like a second skin. The fit is amazing and it's neither too tight nor too loose, just right. I've worn it many times, and it still looks like new. Definitely worth every penny!* Another review of a T-shirt:*"	3.848
"Here is a review of a T-shirt: *I regret buying this T-shirt. The fabric is stiff and uncomfortable, and it doesn't breathe well, making it unsuitable for warmer weather. The stitching started coming apart after a couple of washes, and it just looks cheap overall. I would not recommend this shirt to anyone.* Another review of a T-shirt: *"	5.412
"Here is a review of a T-shirt: *I had high hopes for this T-shirt, but it fell short in every way. The material feels low-quality and rough, and the seams are already coming undone. It also shrank significantly after the first wash, making it unwearable. Save your money and avoid this one.* Another review of a T-shirt: *"	3.077

Note: The asterisks (*) denote line breaks. Perplexity refers to the average perplexity across a sample of 5 synthetic product reviews generated with the given prompt.

While we found that the synthetic product reviews generated by LLaMA 3 performed well with respect to perplexity (see Table I and Table II), they generally exhibited less variation and, at times, more hallucinations. This issue is further discussed in the conclusion of our paper.

[4]The asterisks (*) denote line breaks.

Finally, we generated a sample of 10,000 synthetic product reviews per LM, each containing between 50 and 200 tokens. This range was chosen to balance the reduction of hallucinations, manage computational costs, and ensure sufficient content for the manual component of the market analysis.

A.3 The Decoding Algorithm for Text Generation – Nucleus Sampling

Nucleus sampling (also top-p sampling), introduced by Holtzman et al. (2020), is a stochastic method, where at each step of the text generation, the algorithm selects the next token from the smallest set of tokens whose cumulative probability meets or exceeds a threshold p. Mathematically this can be expressed as

$$\sum_{x \in V^{(p)}} P(w_i|w_1, ..., w_{i-1}) \geq p,$$

where $V^{(p)} \subset V$ is the top-p vocabulary and $p \in [0, 1]$ is the pre-determined threshold. This threshold is usually set at $0.7 \leq p \leq 0.95$, as this generates consistent good quality text (DeLucia et al., 2021; Holtzman et al., 2020).

Let $Z = \sum_{x \in V^{(p)}} P(w_i|w_1, ..., w_{i-1})$. The next word is then sampled from a re-scaled version of the original probability distribution:

$$P^*(w_i|w_1, ..., w_{i-1}) = \begin{cases} P(w_i|w_1, ..., w_{i-1})/Z & \text{if } w_i \in V^{(p)}, \\ 0 & \text{otherwise.} \end{cases}$$

Given this formula, the probabilities of all words that are not in the top-p vocabulary at a given time will be set to 0. Moreover, P^* will be different at each step of the text generation.

Finally, we chose nucleus sampling because other common decoding algorithms, such as greedy search, beam search, and top-k sampling have been shown to perform worse in comparison (Holtzman et al., 2020).[5]

A.4 Perplexity

Perplexity (PPL) is the exponential of the cross-entropy of a probability distribution (Jelinek et al., 1977). It is a metric that measures how perplexed a language model is when generating text. Mathematically, the perplexity of a sentence s with t words $w_i, ..., w_t$ can be defined as

$$PPL(s) = \left(\prod_{i=1}^{t} \frac{1}{P(w_i|w_1, ..., w_{i-1})} \right)^{(1/t)} = \exp\left(-\frac{1}{t} \sum_{i=1}^{t} \log P(w_i|w_1, ..., w_{i-1}) \right),$$

where log denotes the natural logarithm and $P(w_i|w_1, ..., w_{i-1})$ is the probability of a word w_i given the preceding words $w_1, ..., w_{i-1}$ (Aggarwal, 2022, p. 311).

[5] We are aware that newer decoding algorithms, such as mirostat (Basu et al., 2021) η-sampling (Hewitt et al., 2022), and locally typical sampling (Meister et al., 2023), which try to balance quality and diversity of the generated text, exist.

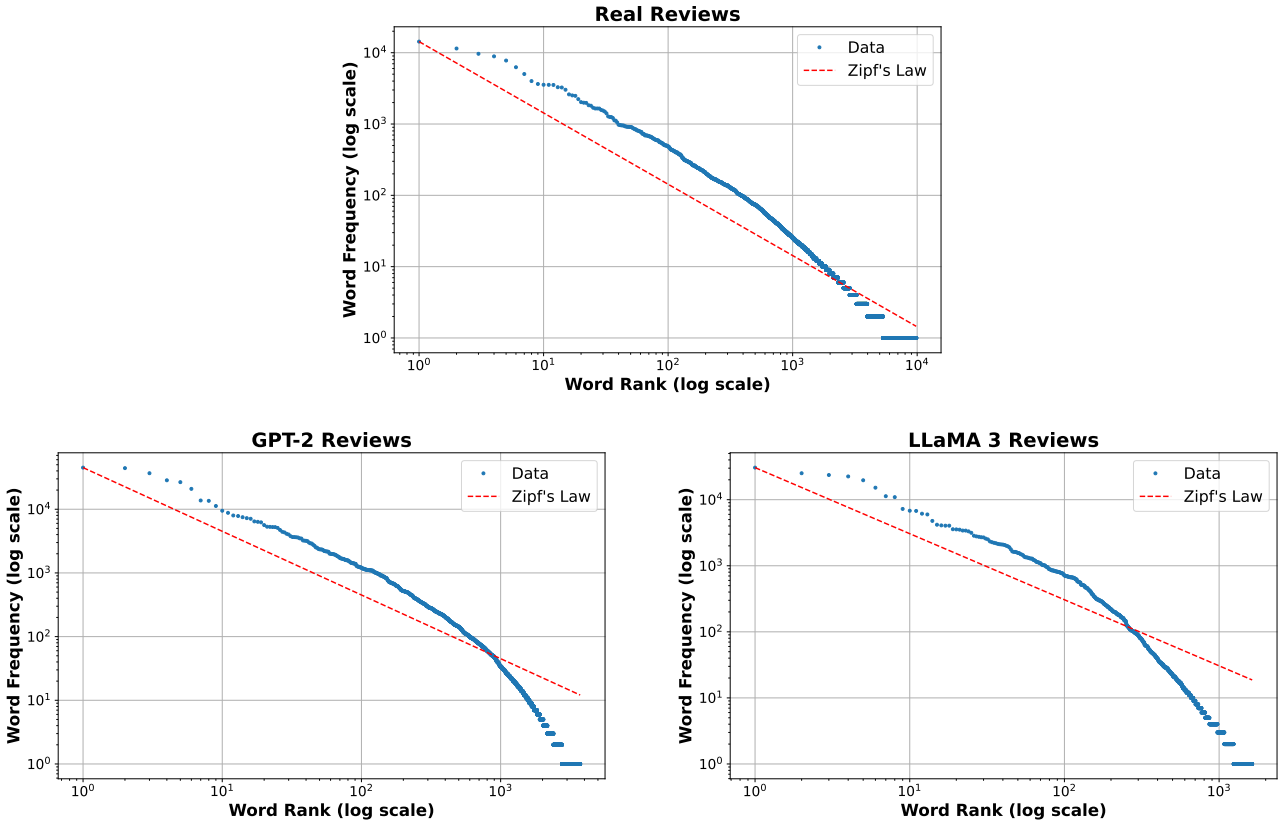

Figure II: Log-Log Plot of Word Rank vs. Word Frequency for each Review Type

A.5 Model Evaluation – Additional Metrics

A.5.1 Statistical Tendencies

The statistical tendencies of a text can be defined as "the degree to which the linguistic distributions of text generated from language models match—or differ from—those of natural language" (Meister & Cotterell, 2021, p. 5330). We focused on Zipf's law, which states that the frequency of a word in a corpus is inversely proportional to its rank in the frequency table (Zipf, 1935, 1949). Mathematically, Zipf's law can be written as:

$$f(w_k) \propto k^{-s},$$

where $f(\cdot)$ denotes the frequency, w_k is the k^{th} word, and s is a normalizing constant. Zipf's law can also be expressed as the probability $P_{zipf}(\cdot)$ of a word w of rank r:

$$P_{zipf}(w_k) = \frac{1}{\zeta(s)} k^{-s},$$

where $\zeta(s) = 1/\sum_{k=1}^{\infty} k^{-s}$ (Meister & Cotterell, 2021, p. 5329).

To assess this relationship, we 1) calculated the rank-frequency of each word, 2) plotted a rank vs. frequency log-log graph, 3) determined the normalizing constant s using the Kolmogorov–Smirnov test (Kolomogorov, 1933; Smirnov, 1948) and finally 4) performed a likelihood ratio test (LRT; Neyman & Pearson, 1928, 1933) to evaluate the goodness-of-fit (cf. Clauset et al., 2009). Figure II illustrates the log-log plot of a word's rank vs. the word's frequency.

Finally, the likelihood ratio test yielded a test statistic of $\chi^2 = 5661.33$ for GPT-2 and $\chi^2 = 2939.25$ for LLaMA 3 (with $s = 0.001$).[6] The p-values associated with this test statistic are $p = 0.0000$ for both PLMs, indicating statistical significance at the 5%-level. Substantively, this result indicates that the product reviews generated by GPT-2 and LLaMA 3 follow a word frequency distribution that is more consistent with Zipf's law than with a uniform distribution, leading to the rejection of the null hypothesis that the data follows a uniform distribution. In other words, the result suggests that both samples adhere to the frequency distribution seen in natural language. However, while the likelihood ratio tests suggest that Zipf's law provides a better fit, the deviations observed in the log-log plots of frequency versus rank (see Figure II) suggest that the alignment with Zipf's law is not perfect. These deviations indicate that both samples of synthetic product reviews do not fully conform to the theoretical predictions of Zipf's law.

A.5.2 Linguistic Features

Linguistic features are the structural aspects of language. We focused on five linguistic features. First, we evaluated the raw text properties of our synthetic reviews, using the average length and standard deviation of word and sentence length. Second, we examine vocabulary richness, which is measured using the type-token ratio.[7] This ratio is calculated by dividing the number of unique tokens (types) by the total number of tokens (types + tokens). Third, we analyze part-of-speech (POS) usage. This involves examining the frequency and patterns of different parts of speech within the text, such as verb usage and noun usage. Finally, we assess readability using the Flesch Reading Ease Score (FRE; Flesch, 1948) and the Dale-Chall Readability Score (DCR; Dale & Chall, 1948). We can compute FRE using Flesch (1948, p. 229)'s formula:

$$\text{FRE} = 206.835 - 846 \times \left(\frac{\text{number of syllables}}{100}\right) - 1.015 \times \left(\frac{\text{sum of all sentence lengths}}{\text{number of sentences}}\right),$$

where $\text{FRE} \in [0, 100]$. DCR can be computed using Dale and Chall (1948, p. 18)'s formula:

$$\text{DCR} = 3.6365 + 0.1579 \times \left(\frac{\text{number of difficult words}}{\text{number of words}}\right) + 0.0496 \times \left(\frac{\text{sum of all sentence lengths}}{\text{number of sentences}}\right),$$

where $\text{DCR} \in [0, 10]$. The results of the linguistic features metrics for our synthetic product reviews and our real product reviews are shown in Table III.

The linguistic features metrics in Table III compare the performance of GPT-2 Medium, LLaMA 3 8B, and a Baseline (real product reviews). The average length of reviews generated by GPT-2 Medium (73.29 words) is substantially higher than both LLaMA 3 8B (41.30 words) and the Baseline (29.30 words), with GPT-2 Medium also exhibiting a greater standard deviation in length (29.87 vs. 9.07 for LLaMA 3 8B).

[6] The optimal value for s was determined by using maximum likelihood estimation (MLE).
[7] The type-token ratio is also known as Heaps' law (Herdan, 1960).

Table III: Results of the Linguistic Features Metrics

	Avg. length	**SD of length**	**TTR**	**% of Verbs**	**% of Nouns**	**FRE**	**DCR**
GPT-2 Medium	73.29	29.87	0.0051	15.09	10.98	93.03	0.66
LLaMA 3 8B	41.30	9.07	0.0040	17.46	10.35	77.33	0.58
Baseline	29.30	31.36	0.0335	17.27	13.17	84.37	1.08

Note: SD = standard deviation; TTR = type-token ratio; FRE = Flesch Reading Ease Score; DCR = Dale-Chall Readability Score. Baseline refers to the values for the real product reviews.

Interestingly, despite the variance in length, the type-token ratio (TTR), a measure of lexical diversity, is significantly lower for both GPT-2 Medium (0.0051) and LLaMA 3 8B (0.0040) compared to the Baseline (0.0335), indicating reduced lexical variety in the generated texts.

In terms of grammatical composition, LLaMA 3 8B produces a higher percentage of verbs (17.46%) than both GPT-2 Medium (15.09%) and the Baseline (17.27%), while GPT-2 Medium shows a slightly higher percentage of nouns (10.98%) than LLaMA 3 8B (10.35%). The Baseline reviews contain a relatively higher noun percentage (13.17%).

Readability metrics further distinguish the models. GPT-2 Medium achieves a higher Flesch Reading Ease (FRE) score (93.03), indicating greater readability, compared to LLaMA 3 8B (77.33) and the Baseline (84.37). However, the Dale-Chall Readability Score (DCR) reveals a slightly different pattern, with GPT-2 Medium (0.66) and LLaMA 3 8B (0.58) producing more accessible text than the Baseline (1.08), suggesting that the generated reviews, despite their differences in length and lexical diversity, tend to be simpler and more easily understood than real product reviews.

Overall, the product reviews generated by both GPT-2 Medium and LLaMA 3 8B exhibit substantial deviations from the Baseline (real product reviews), particularly in terms of lexical diversity and readability. While both models generate text that is simpler and more accessible, this comes at the cost of reduced variety in vocabulary compared to real product reviews.

A.5.3 MAUVE

MAUVE, introduced in Pillutla et al. (2021) and further explained in Pillutla et al. (2023), is a quality measure for open-ended text generation, which compares a generative model's distribution Q with the target distribution P of the real data. Doing so requires considering two types of errors. Type I errors are false positives, where Q assigns high probabilities to text sequences, which are unlikely under P. In other words, a model produces unrealistic, repetitive, or degenerate text that is unlikely to be written by humans. Type II errors are false negatives, where Q cannot generate text, which is likely under P. In other words, the model cannot produce diverse text. These errors can be formalized using the Kullback-Leibler (KL) divergences (Kullback & Leibler, 1951) $\mathrm{KL}(Q|P)$ and $\mathrm{KL}(P|Q)$, which are defined as

$$\mathrm{KL}(Q|P) = \sum_x Q(x)\log\frac{Q(x)}{P(x)} \quad \text{and} \quad \mathrm{KL}(P|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}.$$

However, if the supports of P or Q are not identical (e.g., $P(x) > 0$ and $Q(x) = 0$), one or both KL divergences $\mathrm{KL}(Q|P)$ and $\mathrm{KL}(P|Q)$ will be infinite. Pillutla et al. (2021) overcome this issue by *softly*

measuring the two errors using the mixture distribution $R_\lambda = \lambda P + (1-\lambda)Q$, where $\lambda \in (0,1)$.[8] We can then re-define the (soft) Type I error as $\mathrm{KL}(Q|R_\lambda)$ and analogously the (soft) Type II error as $\mathrm{KL}(R_\lambda|Q)$, which yields the following *divergence curve*

$$\mathcal{C}(P,Q) = \{(\exp(-c\mathrm{KL}(Q|R_\lambda), \exp(-c\mathrm{KL}(R_\lambda|Q))) \ : \ \lambda \in (0,1)\},$$

where $c > 0$ is a hyperparameter for scaling. $\mathrm{MAUVE}(P,Q)$ is then the area under the divergence curve $\mathcal{C}(P,Q)$ and it lies in $(0,1]$, formally written as

$$\mathrm{MAUVE}(P,Q) = \mathrm{AUC}(\{(\exp(-c\mathrm{KL}(Q|R_\lambda), \exp(-c\mathrm{KL}(R_\lambda|Q))) \ : \ \lambda \in (0,1)\}),$$

Using the above formula, we obtained a MAUVE score of 0.78 for GPT-2 and a score of 0.52 for LLaMA 3. These scores can be interpreted as measures of how closely the generated text matches the distribution of real human text, with higher scores indicating a closer resemblance to human-like text (here: real product reviews). A score of 0.78 for GPT-2 suggests a relatively strong alignment with human text, indicating that its generated content is quite similar to real product reviews. In contrast, LLaMA 3's score of 0.52 is moderately lower, suggesting a weaker, but still present, alignment with human-like text. While GPT-2's score can be considered relatively high, LLaMA 3's score is comparatively lower, indicating that it generates text that deviates more from the patterns seen in real reviews. An explanation for the significant difference between the two MAUVE scores is the reduced variability in content in the product reviews generated by LLaMA 3. This lower variability likely causes the LLaMA 3 product reviews to deviate more from the diverse linguistic patterns typically found in real product reviews, leading to a lower score.

A.6 Guidelines for the Human Evaluation of Synthetic Product Reviews

Human evaluation refers to the systematic manual evaluation of LM-generated text by one or multiple human raters. For our evaluation, we used four criteria, each measured on a 5-point scale similar to a 5-point Likert scale (cf. Likert, 1932). This scale ranges from 0 ("strongly disagree") to 5 ("strongly agree"), with a midpoint at 3 ("neither agree nor disagree"). In general, we followed common practice and suggestions from literature reviews (Celikyilmaz et al., 2020; Gehrmann et al., 2023; Howcroft et al., 2020; Novikova et al., 2018; van der Lee et al., 2021). Table IV contains the guidelines for our three criteria that we used for the human evaluation.

[8] A mixture distribution is the probability distribution of a random variable that is derived from a collection of distributions of other random variables.

Table IV: Guidelines for the Human Evaluation

Usability	The review is... **(1)** not about a T-shirt / unknown what it is about / mentioning multiple clothing items & it is very inconsistent with its message (many contradictions in preferences and issues) **(2)** not about a T-shirt / unknown what it is about / mentioning multiple clothing items & it is somewhat inconsistent with its message (only some contradictions in preferences and issues) **(3)** about a T-shirt or T-shirt adjacent article (shirt, sweater, top, jersey) & it is mostly consistent with its message (almost no or no contradictions in preferences and issues) **(4)** about a T-shirt or T-shirt adjacent article (shirt, sweater, top, jersey) & it is consistent with its message (no contradictions in preferences and issues) & it only includes some irrelevant information (content not relevant to T-shirts) **(5)** about a T-shirt or T-shirt adjacent article (shirt, sweater, top, jersey) & it is very consistent with its message (no contradictions in preferences and issues) & it only includes information about the T-shirt
Readability	The review is... **(1)** not understandable even after re-reading it (important: only choose this option, if really nothing is understandable) **(2)** partially understandable but some parts of the review are not understandable even after re-reading it (important: choose this option if there are some or even many things that you do not understand after re-reading) **(3)** understandable but only after re-reading it **(4)** readable and understandable without re-reading it & it has some (visible) mistakes (e.g., grammar or spelling) **(5)** clearly readable and understandable without re-reading & it has no (visible) mistakes
Naturalness	The review... **(1)** does not sound natural at all & it could not have been written by a real person **(2)** contains only some passages do not sound natural **(3)** 's naturalness is not determinable (neither 2 nor 4) (example: long review that is way too over the top) **(4)** sounds natural & some passages sound strange (e.g., it is over the top) **(5)** sounds very natural, virtually indistinguishable from a review written by a real person

B Tables

Table A1: Codebook for QCA

0 Preferences	**1 Issues**
01 Quality → *positive mentions about the quality* **01 01** Seams → *well-made* **01 02** Print → *nice / high quality*	**11 Quality** → *negative mentions about the quality* **11 01** Seams → *poor, stitching bad* **11 02** Print → *low quality, glitter* **11 03** Shape → *difficulties to maintain the shape / stretches out*
02 Material / Fabric **02 01** Thin **02 02** Warm **02 03** Soft **02 04** Not see through **02 05** High quality **02 06** Lightweight **02 07** Breathable **02 08** Thick **02 09** Feels great against skin **02 10** Stretchy	**12 Material / Fabric** **12 01** Thin **12 02** Warm (too warm) **12 03** Soft (not soft) **12 04** See through **12 05** Cheap **12 06** Shiny (too shiny) **12 07** Polyester / Spandex **12 08** Thick (too thick) **12 09** Feels not great → *stiff / rough / itchy / scratchy* **12 10** Stretchy
03 Fit → *positive things like great fit*	**13 Fit** → *bad fit, weird neckline*
04 Sizing → *size is as described / size is accurate*	**14 Sizing** → *inaccurate sizing / too tight*
05 Price → *cheap / good price / affordable*	**15 Price** → *not worth the money*
06 Style / Design → *unique / elegant / timeless*	**16 Style / Design** → *disappointed / cheap*
07 Comfort → *snug / comfortable*	**17 Comfort** → *not comfortable*
08 Delivery / Service → *early / on time*	**18 Delivery / Service** → *late, customer service, stain*
09 Color → *good/ true/ vibrant*	**19 Color** → *wrong*
010 Description online → *looks like photos*	**110 Description online** → *not what was pictured / described online / things in false place*
011 Easy to care → *no ironing, machine washable*	**111 Complicated to care**

Table A1 – continued

0 Preferences	**1 Issues**
012 Washing behavior → *no shrinking / no pulling up / holds shape*	**112 Washing behavior** → *shrunk*
013 Length → *good* **013 01** Sleeves	**113 Length** → *too short / long* **113 01** Sleeves → *too short / long*
014 Zipper	**114 Zipper** → *no zipper*
015 Pockets	**115 Pockets** → *no pockets*
016 Social Company → *supports a good cause*	**116 China** → *from China*

Table A2: Topic Modeling Results for Preliminary Data Analysis

Topic	**Number of Occurrences**	**Keywords**
1	63,886	size, fit, like, small
2	21,696	quality, broke, one, cheap
3	3,919	glasses, sunglasses, lenses, pair
4	3,078	color, colors, green, white
5	1,062	one, star, stars, received
6	1,436	support, arch, shoes, feet
7	1,123	expected, exactly, ok, perfect
8	468	umbrella, water, rain, waterproof
9	1,505	product, watch, work, works
10	1,048	warm, cold, winter, weather
11	287	pay, money, get, waste
12	179	instructions, directions, easy, use
13	98	football, draft, game, love
14	215	bottle, product, water, bottles

Table A3: Hyperparameters for the Fine-Tuning of GPT-2 Medium

Hyperparameter	**GPT-2 Medium**
Learning rate	5e-4
Batch size	2
AdamW ϵ	1e-8
Epochs	5

Table A4: Human Evaluation of the Synthetic Product Reviews

	GPT-2				LLaMA 3			
Criterion	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
Usability	2.89	2.35	2.94	2.53	4.81	4.77	4.75	4.65
Readability	2.64	2.70	2.90	2.66	4.97	4.98	4.96	4.96
Naturalness	2.00	1.97	2.43	1.62	4.33	4.42	4.29	4.23

Table A5: Topic Modeling Results for Market Analysis

Review Type	**Topic**	**Number of Occurrences**	**Keywords**
Real Reviews	1	5,695	fit, size, like, dress
	2	1,832	watch, band, one, earrings
	3	956	quality, product, great, cheap
	4	411	hat, mask, head, nose
	5	400	bag, purse, wallet, zipper
	6	435	color, colors, picture, like
	7	96	see, thin, hole, holes
	8	134	works, work, use, one
	9	41	captive, toxicity, beed, ten
GPT-2 Reviews	1	7,341	shirt, fit, size, like
	2	1,587	soft, warm, material, shirt
	3	327	zipper, one, quality, product
	4	284	wash, washed, shirt, washing
	5	122	hat, head, product, good
	6	115	shirt, stretchy, stretch, material
	7	121	product, item, quality, one
	8	56	long, length, short, color
	9	47	3x, 3xl, says, ordered
LLaMA 3 Reviews	1	5,618	shirt, comfortable, soft, love
	2	2,519	review, reviews, product, shirt
	3	1,554	already, shirt, material, disappointment
	4	127	sentiment, analysis, text, reviews
	5	73	persuasive, writing, essay, reader
	6	29	voice, passive, subject, sentence
	7	29	book, movie, story, restaurant
	8	31	template, structure, reviews, templates
	9	20	best, shirt, water, cold

C Figures

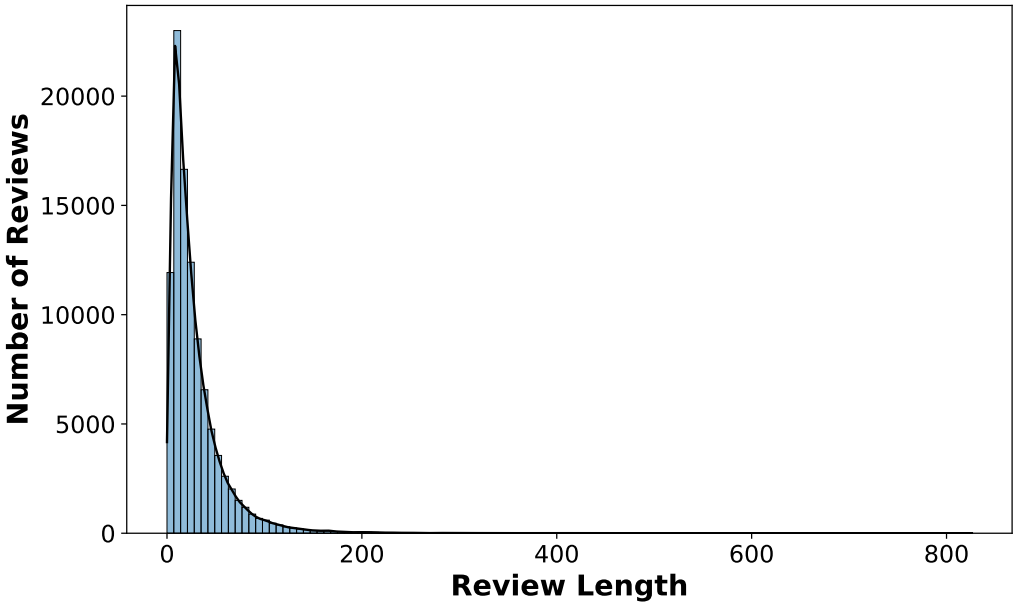

Figure A1: Distribution of the Average Length of the Real Reviews

Figure A2: Word Clouds by Sentiment based on TF-IDF

Figure A3: Word Clouds by Rating based on TF-IDF

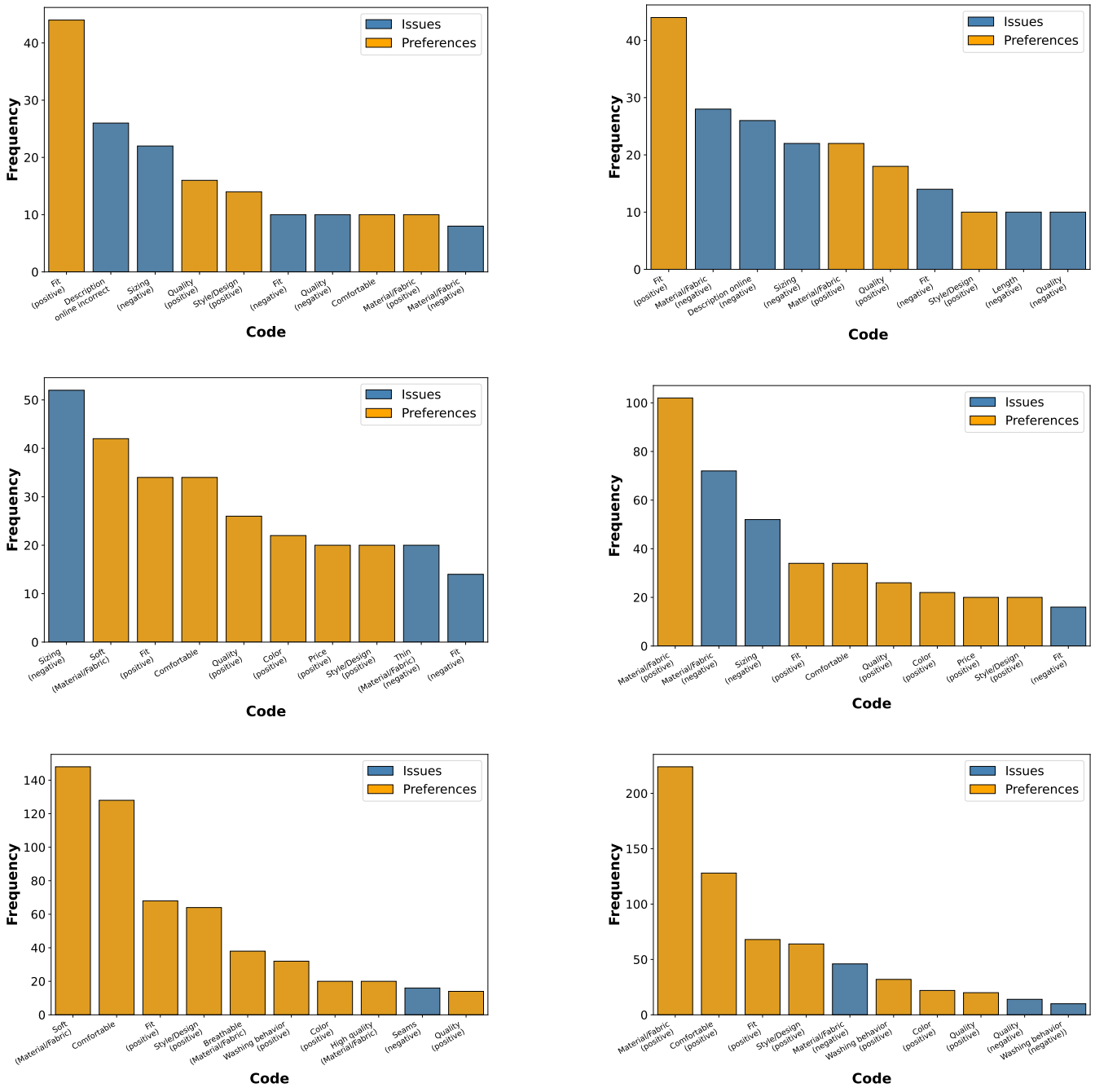

Figure A4: Distribution of Codes from the QCA for the Real Reviews (top), GPT-2 Reviews (middle), and LLaMA 3 Reviews (bottom). Note: The left side shows all codes and the right side shows main codes.

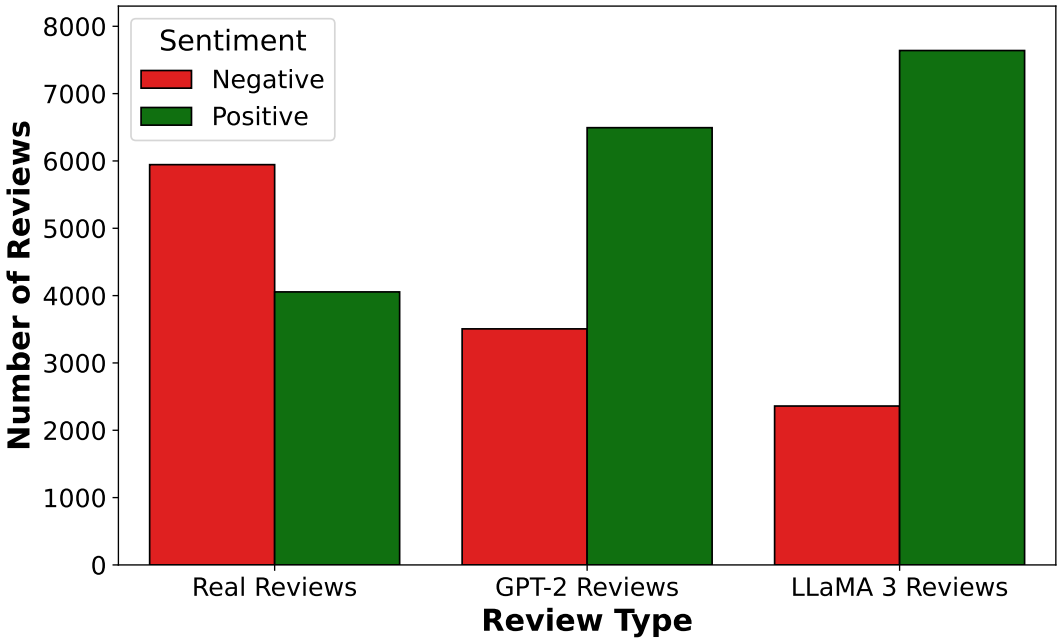

Figure A5: Sentiment Distribution by Review Type

References

Aggarwal, C. C. (2022). *Machine Learning for Text* (2nd ed.). Springer Nature.

Basu, S., Ramachandran, G. S., Keskar, N. S., & Varshney, L. R. (2021). Mirostat: A Neural Text Decoding Algorithm That Directly Controls Perplexity. *International Conference on Learning Representations 2021*, 4816–4828. https://openreview.net/forum?id=W1G1JZEIy5_

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137–1155. https://doi.org/https://www.jmlr.org/papers/v3/bengio03a.html

Boucher, J., & Osgood, C. E. (1969). The Pollyanna Hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *8*(1), 1–8. https://doi.org/10.1016/S0022-5371(69)80002-2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 1877–1901. https://dl.acm.org/doi/abs/10.5555/3495724.3495883

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2023). Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. *arXiv*. https://doi.org/10.48550/arXiv.2312.16171

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of Text Generation: A Survey. *arXiv*. https://doi.org/10.48550/arXiv.2006.14799

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11–20.

DeLucia, A., Mueller, A., Li, X. L., & Sedoc, J. (2021). Decoding Methods for Neural Narrative Generation. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, 166–185. https://doi.org/10.18653/v1/2021.gem-1.16

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F., & Danforth, C. M. (2015). Human language reveals a universal positivity bias. *PNAS*, *112*(8), 2389–2394. https://doi.org/10.1073/pnas.1411678112

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., ... Zhao, Z. (2024). The Llama 3 Herd of Models. *arXiv*. https://doi.org/10.48550/arXiv.2407.21783

Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Gehrmann, S., Clark, E., & Sellam, T. (2023). Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, *77*, 103–166. https://doi.org/10.1613/jair.1.13715

Herdan, G. (1960). *Type–Token Mathematics: A Textbook of Mathematical Linguistics*. Mouton.

Hewitt, J., Manning, C., & Liang, P. (2022). Truncation Sampling as Language Model Desmoothing. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3414–3427. https://doi.org/10.18653/v1/2022.findings-emnlp.249

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The Curious Case of Neural Text Degeneration. *International Conference on Learning Representations 2020*. https://openreview.net/forum?id=rygGQyrFvH

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., & Rieser, V. (2020). Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. *Proceedings of the 13th International Conference on Natural Language Generation*, 169–182. https://doi.org/10.18653/v1/2020.inlg-1.23

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity – a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, *62*(S1), S63. https://doi.org/10.1121/1.2016299

Kolomogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, *4*(1), 83–91.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 9459–9474. https://dl.acm.org/doi/abs/10.5555/3495724.3496517

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *22*(140), 5–55.

Lo, L. S. (2023). The Art and Science of Prompt Engineering: A New Literacy in the Information Age. *Internet Reference Services Quarterly*, *27*(4), 203–210. https://doi.org/10.1080/10875301.2023.2227621

Meister, C., & Cotterell, R. (2021). Language Model Evaluation Beyond Perplexity. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5328–5339. https://doi.org/10.18653/v1/2021.acl-long.414

Meister, C., Pimentel, T., Wiher, G., & Cotterell, R. (2023). Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, *11*, 102–121. https://doi.org/10.1162/tacl_a_00536

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *20A*(3-4), 263–294. https://doi.org/10.1093/biomet/20A.3-4.263

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *231*(694-706), 289–338. https://doi.org/10.1098/rsta.1933.0009

Novikova, J., Dušek, O., & Rieser, V. (2018). RankME: Reliable Human Ratings for Natural Language Generation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 72–78. https://doi.org/10.18653/v1/N18-2012

Pillutla, K., Liu, L., Thickstun, J., Welleck, S., Swayamdipta, S., Zellers, R., Oh, S., Choi, Y., & Harchaoui, Z. (2023). MAUVE Scores for Generative Models: Theory and Practice. *Journal of Machine Learning Research*, *24*(356), 1–92. http://jmlr.org/papers/v24/23-0023.html

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., & Harchaoui, Z. (2021). MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 4816–4828. https://openreview.net/forum?id=Tqx7nJp7PR

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training* (tech. rep.). OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, *1*(8), 9.

Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3411763.3451760

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv*. https://doi.org/10.48550/arXiv.2402.07927

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Smirnov, N. (1948). Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281. https://doi.org/10.1214/aoms/1177730256

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*. https://doi.org/10.48550/arXiv.2307.09288

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech Language*, *67*, 101151. https://doi.org/10.1016/j.csl.2020.101151

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations 2023*. https://openreview.net/forum?id=1PL1NIMMrw

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned Language Models are Zero-Shot Learners. *International Conference on Learning Representations 2022*. https://openreview.net/forum?id=gEZrGCozdqR

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. *International Conference on Learning Representations 2023*. https://openreview.net/forum?id=WE_vluYUL-X

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley.