From Hidden Neural Representations to Actionable Insight: Identifying Beyond-Semantic Drivers of Review Helpfulness through Mechanistic Interpretability

Abstract

Online reviews are crucial for modern commerce, but identifying what makes reviews engaging remains challenging. Reviews contain layers of meaning, some explicit in *what* the text conveys and others embedded in *how* that text is structured and presented. Traditional approaches examine surface-level features, generate broad topics that lack actionable specificity, or require laborious coding of nuanced features. We develop a novel computational approach that uncovers hidden drivers of review helpfulness by analyzing patterns in how text is represented in neural networks. Using 432,248 restaurant reviews, we identify thousands of human interpretable features and determine which strongly impact helpfulness perceptions. Our analysis reveals previously undetectable textual paralinguistic cues, elements like formatting and structure conveying meaning beyond words, that impact engagement. Experimental validation confirms these features causally influence perceived helpfulness. The findings provide concrete guidance: reviewers should structure content using itemized lists with line-breaks and clearly signal updates or transitions with discourse markers. Unlike existing methods that require pre-specifying features, our approach simultaneously evaluates numerous fine-grained patterns to identify subtle but impactful drivers of helpfulness. This approach offers platforms, businesses, and content creators actionable strategies for identifying and generating engaging reviews.

Keywords: Reviews, User-Generated Content, Engagement, Large Language Models, Electronic Word of Mouth, Text Analysis

As an increasing amount of commerce takes place online, user-generated reviews have played a critical role in shaping the marketplace. Shoppers' purchase decisions are influenced by reviews on Amazon. Diners' expectations are set by reviews on Yelp. And consumers' impressions of brands in general are strongly impacted by electronic word of mouth (eWOM) on social media platforms like TikTok and X. Firms, in turn, have turned to reviews on platforms to understand consumer concerns and shape offerings and messaging to improve business.

As reviews have become more important, they have also become more abundant. In this crowded landscape, some reviews are far more engaging, commanding more attention and influence, than others. For example, on Yelp and Amazon consumers can rate reviews as being "helpful." And on social media, posts about brands can be "liked." In both cases, platforms prioritize reviews based on the engagement they garner. Understanding the drivers of engagement is therefore crucial for platforms prioritizing which reviews to display, content creators seeking visibility, and businesses needing to identify potentially influential feedback.

Despite their importance, technological limitations have made it difficult to sufficiently and efficiently understand online engagement of user reviews and to derive holistic yet concrete guidance for platforms, content creators, and businesses. Much of what is known about engagement with reviews is based on aggregate attributes like the presence of images (Li and Xie 2020) or length (Kim et al. 2006; Mudambi and Schuff 2010), broad constructs like emotionality (Berger and Milkman 2012), or high-level topics generated with methods like Latent Dirichlet Allocation (LDA; e.g., Ahn, Son, and Chung 2021).

While these approaches build a valuable foundation for marketing text analytics (Packard, Moore and Berger 2023), they also highlight critical limitations. First, examining aggregate review-level features like review length enables large-scale analysis but provides only indirect evidence about engagement drivers since they ignore review content. Second, while coding for specific constructs (e.g., valence, brand personality) yields nuanced insights into engagement drivers, the cost and labor intensity of human coding restricts analysis to few constructs. Lastly, while automated dictionary-based methods like LIWC can process reviews quickly, they sacrifice accuracy for scale. All of the approaches described above share a critical limitation: they examine features independently or in small groups rather than comprehensively. This prevents us from understanding how features relate to one another or their relative importance. Further, when analyzing review-level features and specific constructs, researchers must decide in advance which features to examine. These approaches risk missing unexpected review features that impact engagement.

While topic modeling techniques like LDA and clustered embeddings can identify broad themes driving engagement without a priori feature specification, they can lack contextual awareness and operate at a level of abstraction that misses crucial details. For example, LDA's bag-of-words representation removes context by discarding word order, formatting, and punctuation. Clustered embeddings preserve context but, like LDA, produce high-level semantic groupings rather than specific textual patterns. Neither approach detects fine-grained features that might form the basis of actionable advice. For example, both methods typically miss textual paralanguage cues (i.e., use of characters, formatting, and words that create meaning beyond semantic content) which drive review helpfulness (Luangrath, Peck, and Barger 2017; Luangrath, Xu, and Wang 2023). What is missing is a scalable approach that systematically uncovers these interpretable, concrete features that lead to actionable guidance.

In response, we develop a methodological framework that learns human-interpretable features from reviews and provides information on the relative importance of the uncovered textual features based on their impact on review helpfulness. This framework detects a wide range of features spanning semantic content, lexical construction, and structural patterns simultaneously and at scale. Concretely, we use specialized neural networks called sparse autoencoders (SAEs) to extract human-interpretable features from a large language model's (LLM) internal representations, which are inherently opaque. This yields a large set of context-sensitive, fine-grained textual features. We then estimate these features' associations with helpfulness using gradient-boosted decision trees (XGBoost) and quantify their impact using SHAP values (SHapley Additive exPlanations). Using this approach with 432,248 restaurant reviews from Yelp, we uncover 18,305 features with adequate empirical support (activating in at least 1% of reviews) and test them simultaneously, while controlling for other variables, to identify the most impactful drivers of review helpfulness.

This study makes several contributions to research. First, we advance the research of textual analytics by introducing methods from mechanistic interpretability (MI; a subfield of AI research) to extract a wide range of relevant textual features from unstructured data simultaneously at scale, understand their relative importance, interpret them, and validate our interpretations. This approach uncovers context-sensitive, concrete, and interpretable textual features that lead to actionable strategies. Unlike topic modeling, which captures abstract thematic patterns, or human coding and lexicon-based methods, which are labor-intensive and require specification of constructs of interest, our approach is both automated and capable of producing interpretable, actionable insights on a multitude of features. Further, features are examined simultaneously, providing the relative importance of uncovered features.

Second, we contribute to the online review and textual paralanguage literature by identifying two new features that operate at the level of format and layout: formatted lists and structural discourse markers. We find these features strongly predict helpfulness in our database of existing reviews. Using controlled experiments, we validate their causal effects on driving helpfulness. These findings demonstrate that presentation choices matter as much as semantic content. They extend prior work that focuses on sentiment, length, images, and topic summaries. Further, they extend textual paralanguage beyond punctuation, emojis, and capitalization to include structure and segmentation choices (Luangrath, Peck, and Barger 2017; Luangrath, Xu, and Wang 2023).

Lastly, we contribute to the proliferating research on LLMs and Generative AI. While most marketing studies to date have treated LLMs primarily as content generators or classification tools, relying on their outputs to summarize, infer, or simulate consumer text, our approach moves beyond the surface level of generated output. Instead, we probe the latent neural representations inside the model and extract relevant features embedded in these internal layers. This perspective transforms LLMs from black-box text generators into data-rich environments whose internal structures can be mined for substantive marketing insights.

Our findings also offer actionable insights for content creators, business owners, and online platforms. For content creators, we demonstrate that specific formatting choices (i.e., lists, line-breaks, and discourse markers) causally increase review helpfulness, enabling them to communicate more effectively. The insight is transferable to other content marketing contexts, such as detecting, testing, and deploying effective social media advertising, email marketing messages, and customer support communications. For business owners, understanding these beyond-semantic signals provides a new lens for monitoring and interpreting customer feedback, enabling them to better identify reviews that will resonate with prospective customers and to encourage more effective forms of engagement from their patrons. For online platforms, our framework suggests design interventions, such as interface nudges or formatting prompts that guide reviewers toward more helpful communication while simultaneously improving the visibility and curation of high-quality content. Collectively, these implications demonstrate how structural features of text can be deliberately leveraged across the ecosystem of review creation, consumption, and management to strengthen consumer engagement.

Conceptual Background

Online Restaurant Reviews

Platforms like Yelp and Amazon are full of user-generated reviews, and this eWOM plays a key role in shaping consumer judgments and behaviors (Christy and Matthew 2012). For both products and services, reviews increasingly influence which products consumers buy and what services they patronize (Babić Rosario et al. 2016; Floyd et al. 2014; You, Vadakkepatt, and Joshi 2015). In many cases, reviews have become a primary source of information for consumers (King, Racherla and Bush, 2014) who often view them as more trustworthy than marketing communications by firms (Bickart and Schindler 2001; Goldsmith and Horowitz 2006).

Influential reviews play a central role in shaping impressions that consumers have of products and businesses (Chen, Dhanasobhon, and Smith 2008; Zhu et al. 2020). User-generated reviews are particularly helpful for restaurants because they are difficult to assess based on their attributes alone (Huang, Lurie and Mitra 2009). Reviews provide users with valuable contextual information that is difficult to grasp based only on descriptions of the restaurant.

Due to the important role that user-generated reviews play in consumer choices, there has been a substantial body of research exploring the ways that consumers use them. Consumers tend to rely more on reviews for services, including restaurants and hotels, than for goods (Babić

Rosario et al. 2016). This may be because reviews provide rich information about restaurants which helps readers assess how well they match their taste (Zhang and Luo 2023). Importantly, not all reviews are treated equally. For example, consumers differentiate between reviews based on how closely they align with their own taste (Wu et al. 2015). They also respond differently to reviews written by experts or influential reviewers, such as Yelp's 'Elite' reviewers (Luca 2016; Nguyen et al. 2021). Overall, the effective use of reviews helps consumers learn about restaurant quality more quickly than they otherwise would (Fang 2022), ultimately improving welfare by increasing the likelihood that they will spend money at high quality restaurants. The effect is particularly pronounced for tourists and travelers who lack local information.

Reviews can also strongly impact restaurants. Being present on a major online review platform, like Yelp, leads to increases in restaurant revenues (Luca, Nagaraj, and Subramani 2023). Further, the benefits of being on Yelp are highly contingent on user reviews. A half-star higher rating causes restaurants to sell out 19 percentage points more frequently (Anderson and Magruder 2012) while a one-star improvement in rating leads to 5-9% increase in a restaurant's revenues (Luca 2016). The content of reviews, both text and photos, is strongly associated with the survival of restaurants (Zhang and Luo 2023). Indeed, reviews can impact whether restaurants are able to stay in business. A few negative reviews early in the life of a business can be difficult to recover from (Motoyama and Usher 2020). Even for established restaurants, negative reviews can cause persistent brand damage unless issues are promptly addressed in a visible way on the platform (Proserpio and Zervas 2018). A few outdated but prominently displayed reviews can persistently harm restaurants (Baskin 2022). These reviews may persist, particularly if they are engaging, because platforms often sort reviews using engagement metrics like helpfulness. And, failure to address a particularly influential review can, for example, create a cascade of negative eWOM through social media (Herhausen et al. 2019).

The strong influence that restaurant reviews exert on consumers means that platforms that host the reviews, businesses that receive the reviews, and content creators that generate the reviews have a strong interest in gaining a better understanding of reviews.

Engagement With Online Reviews

User engagement, such as clicks, shares, and comments, are important metrics of the influence of UGC on online platforms. For online reviews, they are often measured via explicit expressions like "helpful" votes. When reviews are more engaging, consumers are more likely to read and be influenced by them (Packard and Berger 2017). Platforms frequently use such engagement metrics as a way of prioritizing content. For example, social media platforms prioritize content that gets more likes; product or service-related websites often prioritize reviews deemed as more helpful. At the same time consumers use these expressions of engagement to help them prioritize UGC (Dagogo-Jack and Watson 2025). For example, these markers are used by consumers as a heuristic for content quality. Put simply, understanding what drives engagement has therefore become crucial for multiple stakeholders including platforms, business owners, and content creators in the review ecosystem.

While there is a substantial body of research on reviews and engagement, the unstructured, highly contextual nature of review text makes it difficult to isolate important and actionable drivers. Even where large datasets linking text to engagement metrics like helpfulness exist (e.g., the Yelp Open Dataset), these drivers often lie in latent, context-dependent features that standard approaches struggle to recover. Moreover, because text is flexible, the latent feature space is large, complicating identification of the features that matter most.

This complexity has pushed the literature toward observable proxies amenable to standard text analytics. For example, Mudambi & Schuff (2010) find that review extremity and depth, measured using review star ratings and length, impact helpfulness. And Lee, Trimi, and Yang (2017) find that the presence of more nouns, verbs, and adverbs in reviews is associated with helpfulness. There is, however, a growing body of work that goes beyond proxies with topic models and supervised algorithms trained on human-coded labels. For example, Ghose and Ipeirotis (2011) trained a model that classifies sentences as subjective or objective, finding that the subjectivity of reviews is an important predictor of helpfulness. These approaches can recover coarse latent structure or specific, pre-defined constructs, but they still miss many context-dependent features that drive engagement in practice. Here, we discuss our existing understanding of engagement and the methods used to get to that understanding.

Quantifying observable features

Much of the existing work on engagement with reviews focuses on quantifying observable features detectable using standard computational methods (e.g., bag-of-words, parsing metadata, dictionary-based analysis). These studies have uncovered key drivers including the mere presence of images (Li and Xie, 2020), URLs, and hashtags (Suh et al. 2010), as well as structural features of reviews like their length and the percentage of nouns and verbs (Kim et al. 2006; Mudambi and Schuff 2010). More complex methods involving content recognition algorithms have also found that the presence of faces in pictures and the professionalism of images increases engagement on social media (Li and Xie, 2020).

In many cases, these features are used as distant proxies for psychological processes driving engagement. For example, Berger and Milkman (2012) use emotionality in text as a proxy for physiological arousal and attention to understand the virality of content. Wang et al.

(2021) use vocal indicators of stress and focus as proxies for competence and warmth attributions, finding that these inferred traits drive persuasion success on Kickstarter. And Kanuri, Hughes, and Hodges (2024) use color complexity in Facebook images as a proxy for processing mode shifts, showing that central (versus peripheral) processing drives engagement.

These studies achieve high ecological validity by analyzing large-scale behavioral data directly from the platforms under investigation. However, standard computational methods often miss context-dependent features that could be critical engagement drivers. One important but hard-to-detect feature that exemplifies this challenge is textual paralanguage, the use of characters, formatting, and words that create meaning beyond semantic content (Luangrath, Peck, and Barger 2017), which predicts social media engagement more effectively than standard text analysis alone (Luangrath, Xu, and Wang 2023). Traditional methods struggle to capture it: bag-of-words approaches remove paralinguistic markers (case, spacing, punctuation) as noise, topic models lack the sensitivity to detect such features, and dictionaries miss context-dependent or novel expressions. Luangrath, Xu, and Wang (2023) advanced this area by developing PARA, a rule-based algorithm that successfully identifies pre-defined instances of paralanguage. While PARA provides a valuable foundation, its rule-based approach limits detection of novel or highly contextually sensitive features. Such limitations in detecting context-dependent features like paralanguage highlight gaps in our ability to fully quantify engagement drivers.

Another key limitation of existing methods is that they rarely provide an indication of how important a given feature is compared with other plausible features. Given technical limitations, when studies do look at multiple features together, they only examine a few at a time (e.g., Mudambi & Schuff 2010 examine rating, length, and product type together). Without the ability to compare the relative impact of different engagement drivers, practitioners lack concrete guidance on which features to prioritize. This issue can be mitigated with meta-analysis (e.g., Wang, Wang, and Yao 2019). Even with meta-analysis, these methods cannot find features that are not already predicted by theory.

Feature coding

Another stream of research studies specific constructs using human or algorithmic coders. For example, Yazdani, Chakravarty, and Inman (2025) code emotional expressiveness of people on a crowdfunding website, and Villarroel Ordenes et al (2019) classify advertisements as action- or information-oriented. Human coders provide an unparalleled ability to detect subtle and contextually sensitive features in text. At the same time, because they are slow and expensive, they are typically unsuited for analyzing large databases of reviews. Algorithmic coders, either using dictionaries of words (e.g., LIWC, Pennebaker, Booth, and Francis 2001) or supervised machine learning algorithms trained on human coding (e.g., Villarroel Ordenes et al. 2019) can code large databases. However, these methods, again, require the construction of dictionaries or the collection of human training data, making them practically difficult to apply. And while machine learning algorithms trained on human responses can provide highly context sensitive results, dictionary-based methods like LIWC lack contextual sensitivity, resulting in substantial noise (Hartmann et al. 2019). More recently, synthetic coders that use LLMs to simulate human coders have shown some promise in combining the depth and flexibility of human coders with the scalability of algorithmic coders (e.g., Matter et al. 2024).

These techniques, like observable features methods, test theories in isolation making comparison difficult and the discovery of unexpected drivers unlikely.

Experimental Studies

Some research on engagement has relied on laboratory-based studies to understand the factors driving engagement with UGC. For example, Ceylan, Diehl, and Proserpio (2024) supplement correlational evidence that similarity between text and photos in reviews drives helpfulness with experiments. Yin, Bond, and Zhang (2017) manipulated expressed arousal in reviews to understand its impact on helpfulness. Experimental methods allow researchers to make strong causal claims about factors driving engagement within an experimental setting. But, they are also less ecologically valid than the techniques described above because they often require artificial environments and rely on potentially un-representative samples of experimental participants. Importantly, while field studies can circumvent issues of ecological validity, they are typically expensive and complex to execute. Further, these methods also require pre-specified theories, with the same limitations discussed earlier.

Topic modeling

More recently, several analyses of engagement have focused on more semantically sensitive techniques like analyzing embeddings and topic modeling with LDA. Ahn, Son, and Chung (2021) use LDA to identify topics across five organization types responding to natural disasters, finding that certain topics predict engagement. For instance, media organizations' breaking news tweets with video outperform disaster impact coverage. Cao, Duan, and Gan (2011) use Latent Semantic Analysis (LSA) and Support Vector Machines (SVMs) to predict the helpfulness of reviews, finding that 11 out of 100 empirically derived semantic characteristics predict helpfulness. Importantly, they do not interpret these features. Similarly, Mahdikhani (2022) found that topic modeling using LDA, when combined with either embeddings or Term Frequency-Inverse Document Frequency (TF-IDF) vectors, strongly predicts engagement.

These techniques generate broad and empirically derived semantic topics that are difficult to detect using earlier methods. Further, they allow for the comparison of many topics simultaneously, providing insight into which ones are the most impactful drivers of engagement (e.g., Cao, Duan, and Gan 2011). At the same time, the topics identified by these techniques are typically broad, making it difficult to use them as a basis for concrete guidance to practitioners.

Finding Hidden and Actionable Drivers of Review Engagement

A key technological development in recent years has been the popularization of LLMs. LLMs are a specific form of AI that learns to predict patterns in text through massive-scale training on diverse written content, developing emergent capabilities far beyond their simple next-token prediction objective (Brown et al. 2020; Radford et al. 2021). The most well-known among these are OpenAI's GPT family of models. In business, these models have been adopted widely within organizations (Singla et al. 2024), with 63% of marketers surveyed (SAS 2024) saying that they use LLMs in their professional lives daily.

Most current business uses of LLMs are focused on the outputs they generate. For example, LLMs have been applied in customer service settings (Challa 2025) as well as for automatically managing brand reputation (Brandwatch 2024). They have also been used for automated content creation allowing for hyper-personalization at scale (Harkness et al. 2023). And in software development, the rise of autonomous programming agents and natural language based software engineering platforms have dramatically improved productivity for some organizations (Deniz et al. 2023) and made software development more accessible to people without technical backgrounds (Harkar 2025). These applications leverage LLMs' ability to process and generate contextually nuanced text.

As a result of the growing popularity of using LLM outputs in business settings, research has focused on understanding the capabilities of (e.g., Arora, Chakraborty, and Nishimura 2025; Goli and Singh 2024; Li et al. 2024) and consumer responses to (e.g., Bai et al. 2025; Kreps, McCain, and Brundage 2022; Zhang and Gosline 2023; Jakesch, Hancock, and Naaman 2023) these outputs. A smaller stream of research has also focused on using LLM outputs to aid in business-related research functions. For example, in market research, there has been much interest in "digital twins" which may allow firms to simulate consumers to predict demand (Toubia et al. 2025). Research has used LLMs to conduct sentiment analysis (Krugmann and Hartmann 2024), predict polarization of media content (Yoganarasimhan and Iakovetskaia 2024), identify a company's marketing mix from social media (Ringel 2023), and understand customer experience to infer customer needs (Timoshenko, Mao, and Hauser 2025).

These applications treat LLMs as output generators, overlooking direct uses for their internal representations. Prior to generating text, LLMs create numerical representations capturing thousands of semantic, lexical, and structural patterns. When given a review, an LLM encodes thousands of relevant features internally but outputs only a few words or sentences. This is analogous to how human speech captures only a fraction of a speaker's mental activity. But unlike human thoughts, LLM representations can be practically extracted and analyzed at scale, providing rich, untapped information about reviews. Businesses can use these representations to analyze millions of posts to identify engagement drivers, optimize algorithms, and refine messaging.

Extracting actionable insights from these representations, however, has been challenging. For example, BERT, an early LLM (Devlin et al. 2019), generates 768-dimensional representations that researchers clustered to create semantic categories (Grootendorst 2022). But because these vectors encode many concepts simultaneously, the resulting categories tend to be broad (e.g., reviews highlighting authenticity), offering limited practical guidance.

Recent work in MI, however, has created tools which allow for these model internal representations to be decomposed into a large number of specific, human-interpretable features (Bricken et al. 2023). MI research studies the way that neural networks process information. For example, Templeton et al. (2024) find that Anthropic's Claude 3 Sonnet model forms interpretable representations of features like transit infrastructure (e.g., bridges, tunnels, and aqueducts) and monuments (e.g., pyramids, the Alamo, the Louvre). In general, this research has either focused on understanding how LLMs work (e.g., Cunningham et al. 2023) or ensuring that LLMs do not produce dangerous outputs (e.g., Templeton et al. 2024). However, the techniques developed in MI also offer marketers an unprecedented ability to understand large databases of unstructured data using LLM representations. Existing MI work rarely links these representations to consumer outcomes. In this article, we adapt MI tools to build a framework that maps specific, human-interpretable textual features to review helpfulness, offering practical strategies for finding and generating engaging reviews.

Methodological Framework

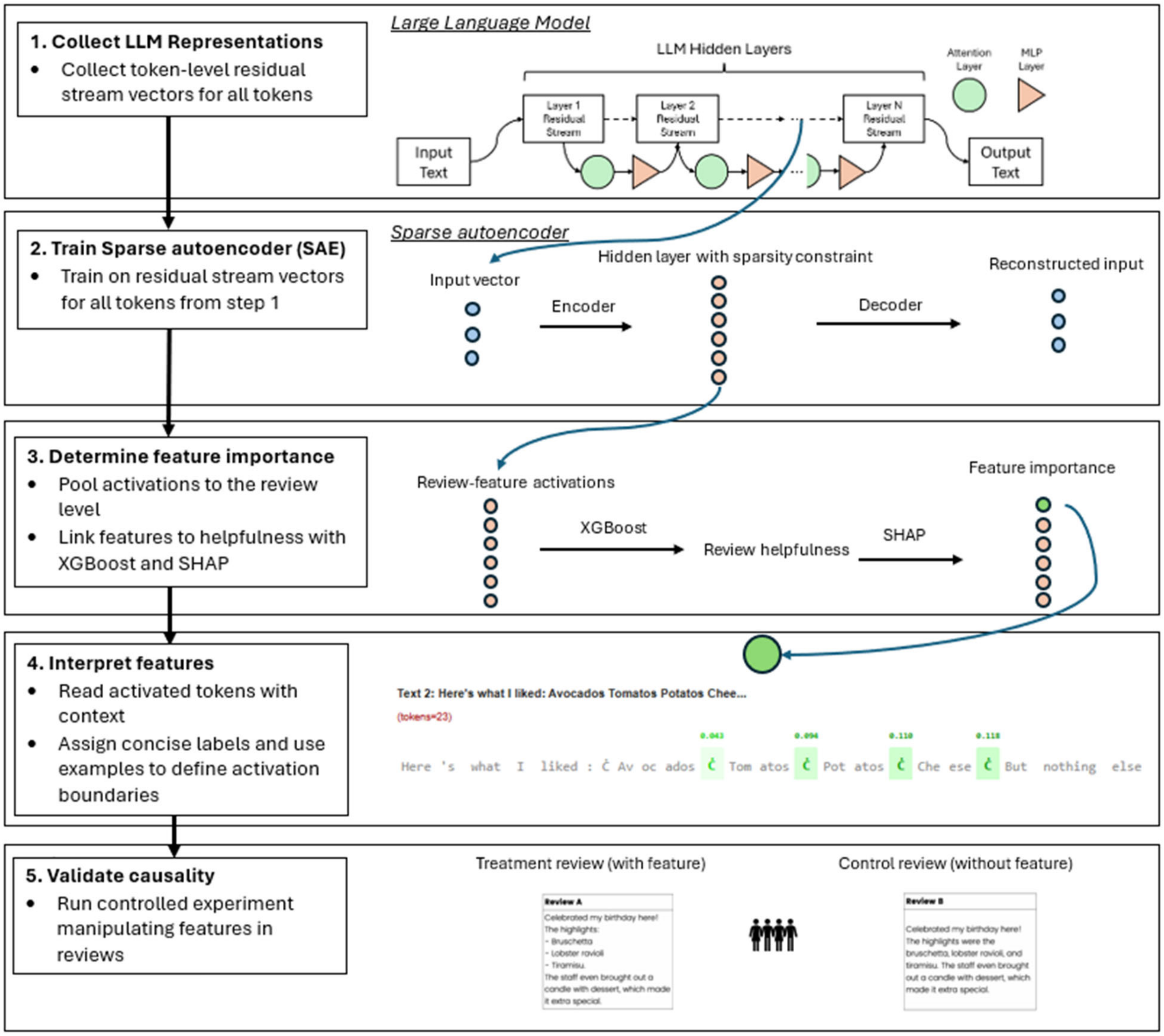
Our framework consists of five steps (Figure 1). We first collect the LLM's rich internal representations for all tokens in our data. We then use these representations to train a simpler, secondary neural network called an SAE. Unlike existing methods, this approach generates numerous specific and human-interpretable features that span semantic, tonal, rhetorical, and structural dimensions of reviews. We then aggregate the activations of all features at the review level and identify key drivers of review helpfulness. We interpret the meanings of impactful features by examining the feature-activating tokens and their context and use LLMs to define their activation boundaries. We illustrate interpretation with two specific textual paralanguage features which were understudied due to their contextual sensitivity. Finally, we validate the impact of the textual features using controlled experiments. Taken together, our approach applies methods from MI to systematically uncover previously inaccessible features that provide platforms, businesses, and content creators with precise, actionable guidance for identifying and generating impactful reviews.

Collecting LLM Activations

Deep neural networks are composed of many hidden layers. Large vectors of information are passed between these layers (LeCun, Bengio and Hinton, 2015; Goodfellow, Bengio, and Courville, 2016). Unlike with traditional neural networks where output vectors from one layer are sent directly to the next layer, with LLMs, outputs from one layer are added to their inputs before being passed to the next layer. This stream of vectors is called the residual stream. At each layer, the model reads from the residual streams, processes them, and then adds back to the residual streams for processing in the next layer. This process of adding to the residual stream rather than passing wholly new vectors across layers allows models to maintain and reprioritize context as they process data. Importantly, because of the fixed dimensionality of the residual stream, models condense and highlight information across layers (Vaswani et al. 2017).

LLMs process text as tokens which are common sequences of characters (typically shorter than words) found in text. These tokens exploit statistical relationships in text that allow for a balance between granularity of information processed by the LLM and computational efficiency owing to the model not having to process every character separately (Sennrich, Haddow and Birch, 2015).

Figure 1: Process for Extracting, Interpreting, and Validating Features.



1. Collect LLM Representations
- Collect token-level residual stream vectors for all tokens

2. Train Sparse autoencoder (SAE)
- Train on residual stream vectors for all tokens from step 1

3. Determine feature importance
- Pool activations to the review level
- Link features to helpfulness with XGBoost and SHAP

4. Interpret features
- Read activated tokens with context
- Assign concise labels and use examples to define activation boundaries

5. Validate causality
- Run controlled experiment manipulating features in reviews

Review A

Celebrated my birthday here!
The highlights:
- Bruschetta
- Lobster ravioli
- Tiramisu.
The staff even brought out a candle with dessert, which made it extra special.

Review B

Celebrated my birthday here! The highlights were the bruschetta, lobster ravioli, and tiramisu. The staff even brought out a candle with dessert, which made it extra special.

Each token, t, is translated into an embedding x_t using an embedding matrix W_E, such that $x_t = W_E t$. The embedding matrix, W_E, serves as a lookup table trained alongside the main model that translates tokens into continuous vector of numbers, x_t. These vectors, along with positional information of tokens, initializes the residual stream, a d dimensional vector, which is fed to the next part of the LLM. At this point, for each token, the residual stream carries forward information from all previous processing layers, both about the token itself and its preceding context (Vaswani et al. 2017).

Within each layer, tokens are processed simultaneously using a transformer. Transformers are composed of two components, an attention layer, which moves information between tokens' residual streams, followed by a Multi-Layer Perceptron (MLP) layer, which calculates an activation based on the residual stream passed from the attention layer. After all layers, the model uses the residual stream to produce a probability distribution over a vocabulary (Vaswani et al. 2017). Models like GPT-2, which we use in our analysis, are trained to find parameters θ which minimize cross-entropy loss L for predicting the next token based on these probabilities averaged over $T - 1$ predictable positions in the sequence. Specifically,

$$L = -\frac{1}{T - 1}\sum_{t=1}^{T-1} log\; p_{\theta}(x_{t+1}|x_{1:t}),$$

where x_{t+1} is the next token and $p(\cdot\,|x_{1:t})$ is the model's conditional distribution over the next token (Goodfellow et al. 2016).

In most commonly used LLMs, attention is "causally masked" such that only context that comes before the current token can impact the representation of a token in the current layer. This masking does not occur in bi-directional models such as BERT (Devlin et al. 2019). Because of their summary nature, collecting residual streams allows us to better understand individual pieces of text in context.

In our analysis, we use GPT2, a precursor to more recent LLMs from OpenAI (Radford et al. 2019). We use this model for three reasons. First, GPT2 is open-source, meaning that researchers can have full control of the model. Most commonly used models are closed-source (e.g., GPT4o, Claude Opus 4.1, Grok 4, and Gemini 2.5 pro). Open-source models are necessary for our work because closed models do not allow researchers access to the residual stream information which we need for our method (Abdurahman et al. 2025). Second, GPT2 is a relatively light-weight model, making it a practical option for researchers or practitioners. While other open-source models exist (e.g., Llama 3 and 4 class models from Meta), they require substantially more computing power. Finally, GPT2 is commonly used in interpretability research (e.g., Wang et al. 2022; Meng, Andonian and Belinkov, 2022; Rajamanoharan et al. 2024; Gao et al. 2024).

Training SAEs to Detect Human Interpretable Features

While residual streams encode rich, contextually sensitive information, they are dense, in that most dimensions are active (i.e., nonzero). This density makes it difficult to interpret the representation for any given piece of text. We therefore train an SAE to decompose these dense residual stream vectors into a large number of features that activate selectively, enabling interpretation. In effect, the SAEs disentangle the dense vectors into larger sparse ones where each vector component tends to code for a single feature.

An autoencoder (AE) is a neural network trained to take an input and reconstruct it (Goodfellow et al. 2016). They have been applied in marketing research across a broad range of applications including understanding the relationship between features in logos and brand personality (Dew, Ansari, and Toubia 2022), predicting the appeal of new product designs (Burnap, Hauser and Timoshenko 2023), improving collaborative filtering (Sedhain et al 2015), and predicting credit worthiness (Mancisidor et al. 2021).

In our case, the AE creates a reconstruction $\hat{x}$ of the residual stream x for each token. First, the encoder projects the token's residual stream vector $x \in \mathbb{R}^d$ onto m feature directions using weight matrix $W_{Enc} \in \mathbb{R}^{m \times d}$, producing scores h:

$$h = xW_{Enc}^{\top}$$

h is the vector of projection scores of the input x onto each feature direction. It then calculates activations for each feature,

$$z = f(h),$$

where $f(h)$ is typically a non-linear transformation (e.g., rectified linear unit activation function). Finally, the decoder uses weight matrix, $W_{Dec} \in \mathbb{R}^{d \times m}$ to reconstruct the input from the activations z:

$$\hat{x} = W_{Dec} z$$

An SAE is an AE that enforces a sparsity constraint in its hidden layers. For a given token, it is trained to use a small number of its hidden layer neurons. In line with prior MI research (e.g., Cunningham et al. 2024; Gao et al. 2024), we adopt an SAE with a single hidden layer that is multiple times larger than the dimensionality of the residual stream. Combining this large dimensionality with a sparsity constraint on activations forces each neuron within the SAE to specialize, increasing the likelihood of human-interpretable features (Bricken et al. 2023; Cunningham et al. 2024; Gao et al. 2024).

We enforced sparsity with Top-K gating (Makhzani and Frey 2013): for each residual stream vector, x, we keep only the k highest scores, $z = T_k(h)$ from the encoding stage to reconstruct $\hat{x} = W_{Dec} z$. While earlier research into interpretability used an L1 penalty term to enforce sparsity, this led to large numbers of neurons that never activated in the trained model, causing computational inefficiency, instability during model training, and less interpretability for models with a given hidden layer size (Gao et al. 2024). More recent MI work has focused on Top K-gating, which tends to lead to fewer unused neurons (Rajamanoharan et al. 2024).

We train the SAE with back-propagation where error e for each token i is calculated as

$$e_i = x_i - \widehat{x_\iota}$$

where $x_i \in \mathbb{R}^d$ is the original residual stream vector and $\widehat{x_i}$ is the reconstructed vector (Goodfellow et al. 2016). We update parameters using minibatches of size B to speed computation and smooth updates.

For each batch and residual stream of dimension d, this loss, L, is calculated as:

$$L = \frac{1}{Bd}\sum_{i=1}^{B}\sum_{j=1}^{d}(x_{ij} - \hat{x}_{ij})^2$$

Finally, to further minimize the presence of dead neurons, we implement a small auxiliary loss to reactivate neurons that have not fired for many tokens, and to reduce noise in model updates, we implement an exponential moving average model (Gao et al. 2024). We detail these procedures in the Web Appendix.

Identifying SAE Hidden Layer Features Associated with Helpfulness

To identify the features most associated with helpfulness ratings, we first ran the residual stream vector for each token through our trained SAE and collected feature activations. Since helpfulness ratings are at the review level, we then pooled the activations for each review by taking the maximum activation for each feature rather than the average since the sparsity of our activations would push averages to zero.

We then employed gradient boosting regression via the XGBoost algorithm (Chen and Guestrin 2016). XGBoost is well-suited to our task for three reasons. First, it can handle high-dimensional, sparse input data, such as our 24,576 latent text features derived from the SAE, along with control variables capturing restaurant fixed effects, review timing, and star rating. Second, as an ensemble of decision trees, XGBoost captures non-linear relationships and higher-order interactions between features. Third, XGBoost is computationally efficient relative to both traditional bagging methods such as Random Forests and more complex deep learning architectures, as it leverages sparsity-aware algorithms, parallelized tree construction, and optimized memory usage to handle large-scale, high-dimensional datasets with relatively low computational cost (Chen and Guestrin 2016). While XGBoost has been widely adopted in computer science domains, its application in marketing research—particularly for combining unstructured text features with structured control variables—remains limited (Ma and Sun 2020).

To interpret the contribution of each feature to the model's predictions, we applied SHAP (SHapley Additive exPlanations; Lundberg and Lee 2017), which assigns each feature a Shapley value, its marginal contribution to the predicted outcome for each observation. SHAP provides consistent, model-agnostic attributions and can be decomposed to analyze effects when features are "active" (nonzero) versus inactive. This allows us to isolate features whose presence in a review text meaningfully increases or decreases predicted helpfulness, even after controlling for confounding variables. SHAP has been adopted in business and marketing research to provide interpretable insights from complex machine learning models (von Zahn et al. 2025).

Interpreting and Validating Features

We employ a three-stage process to interpret and validate SAE features that predict review helpfulness: (1) feature recognition to determine what textual patterns activate each feature, (2) model-based validation to computationally test feature behavior, and (3) experimental validation to confirm real-world impact. While we recognize many features in stage 1, we focus model-based and experimental validation on paralinguistic cues, since conventional methods have difficulty detecting them.

In the feature recognition stage, we scan the corpus to find the tokens where each SAE feature of interest activates most strongly. This is analogous to how topic modeling methods like LDA or clustered embeddings label topics by looking at the most representative phrases or documents (e.g., Timoshenko and Hauser 2019; Kaul et al. 2025). However, unlike many topic modeling analyses, the granularity of our features allows us to focus on individual tokens and the context prior to them.

We read the activated token together with its left context because SAE activations are derived from an LLM's residual stream, which summarizes only the preceding context. Many activations cannot be meaningfully interpreted in isolation. For example, one feature appeared to activate strongly on line-breaks. Examining the context in which the activated line-breaks occurred, we found that the feature only activated when the line-breaks were parts of formatted lists, but not ordinary paragraph breaks. Based on the highly activated tokens and surrounding context, we identify a single, coherent theme that summarizes when this feature is activated.

In the model-based validation stage, for each target feature, we create phrases which should activate the feature based on the results of the feature-recognition stage. We then pass each phrase through the LLM, collect residual stream vectors, and then feed those residual streams through the SAE to extract feature activations. The target feature should activate strongly at the expected token. Next, we create counter-examples to ensure that we know when the feature does not activate. These typically use the same token in different contexts or maintain structure while changing the words. We also explore the boundary of the feature's meaning by trying partial or alternative expressions. By repeating this computational validation loop, we produce stable feature labels and confirm that they reflect the model's activation behavior.

While the model recognition and model-based validation interpret features, they do not causally link features to perceived helpfulness. We therefore run controlled experiments. For each feature, we create matched pairs of reviews that differ only in the presence of the feature. The treatment version includes the feature. The control version removes it while keeping meaning, length, rating, and wording as close as possible. Participants are randomly assigned to one pair and indicate which review is more helpful.

Implementation

Extracting Human-Interpretable Features

In our analysis we use the Yelp Open Dataset, which includes reviews and metadata for 150,346 businesses across 11 metropolitan areas. We filter for businesses classified as "Mexican," yielding 432,248 reviews from 4,600 unique businesses across 10 metros. We focus on a single cuisine to reduce topical variance and improve feature purity when training a compact SAE.

To collect the residual stream vectors, we submit 51,702,947 tokens extracted from the reviews in non-overlapping 64 token chunks to GPT2. This means that, for a given token, only the tokens that come prior to it within a chunk can influence that token's residual stream. Chunking in this way reduces peak memory use and computation while making it simpler to organize data in downstream operations. Importantly, in practice, our features are typically not influenced by more than a few prior tokens of context.

GPT2 uses a 768-dimensional residual stream and has 12 hidden layers. We extract residual stream data prior to the 8th layer. We use a relatively late layer because, in general, LLM residual streams tend to encode higher level and more abstract information in later layers (Ethayarajh 2019). However, layers near the end (i.e., layer 12) tend to be unsuitable for interpretation because they encode information more relevant to producing responses rather than interpreting incoming information. MI often focuses on residual streams from middle layers (e.g., Meng et al. 2022; Gao et al. 2024). We collect residual stream information from all 51,702,947 tokens from 432,248 reviews resulting in a 51,702,947 X 768 matrix.

In general, SAEs with greater numbers of hidden layer features and sparser activations create higher quality models (Gao et al. 2024). At the same time, there is a tendency for models with too many hidden layer features to generate an excessive number of dead neurons effectively resulting in less sparse models. We therefore tested three specifications of SAEs and ran all tokens through the training loop once.

All models selected only the top 32 neurons but varied in their level of expansion. We used 3 levels of expansion (16, 32, and 64x, with 12,288, 24,576, and 49,152 hidden layer neurons respectively). As expansions get larger, we use lower base learning rates to prevent training failure related to sudden jumps in loss, weights growing without bound, and large numbers of dead neurons, as recommended by Gao et al. (2024). We then use adaptive moment estimation with decoupled weight decay (AdamW; Loshchilov and Hutter 2017) to adapt the learning rate of each parameter separately while slowly pushing them toward zero to increase model stability and help generalization by keeping the model from learning from noise. We train the model using GPT2 activations that have been normalized within each token by subtracting the mean and dividing by the L2 norm so that the vector has a unit length of 1. Normalizing within tokens creates scale invariance across inputs, preventing those with high overall activation levels from dominating the training process (Ba, Kiros, and Hinton 2016).

We settled on the 32x expansion model because it offered a lower normalized reconstruction error (.090) than the 16x expansion (.100) and a lower dead neuron percentage (.43%) than the 64x expansion (1.24%). Notably, the number of dead neurons in the 32x expansion was stable throughout the latter half of training while the number in the 64x expansion was in an upward trajectory, suggesting that the model may not have stabilized.

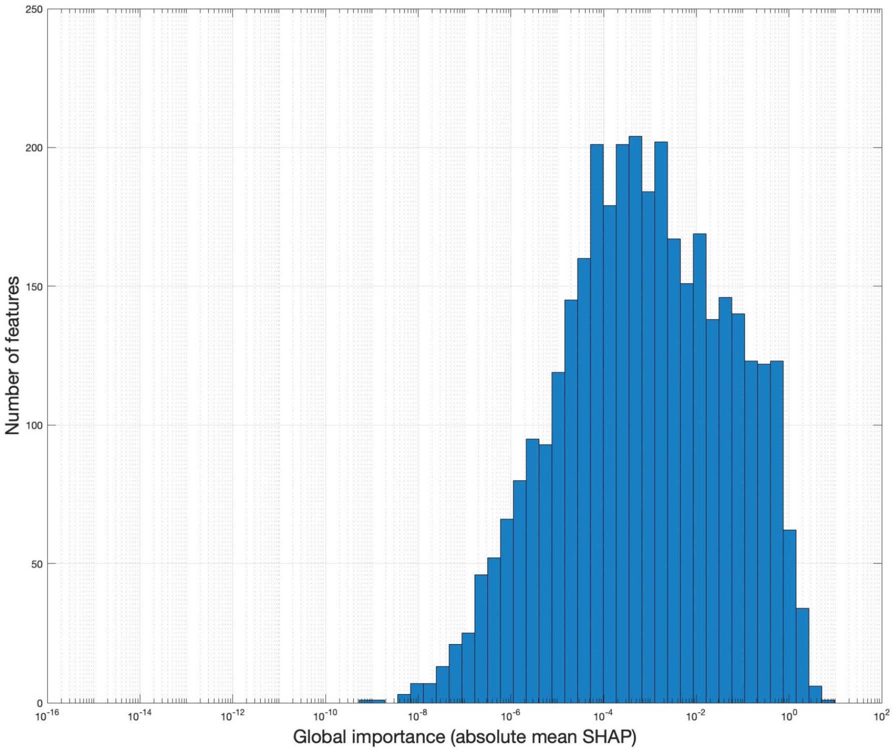
Identifying Features Associated with Review Helpfulness

We focus our feature impact identification on well-supported signals by first selecting the SAE-derived text features that activate in at least 1% of reviews. This 1% activation prevalence filter is analogous to standard document-frequency pruning in text mining, which ensures each reported feature has adequate empirical support and reduces variance from very rare features (Grootendorst 2022; Tirunillai and Tellis 2014). We employed gradient boosting regression via the XGBoost algorithm with the remaining 18305 features. We ran five-fold cross-validation. In each fold, we fit gradient-boosted decision trees that learn to minimize the squared difference between predicted and observed helpful votes, with trees limited to a maximum depth of five and an ensemble size of one hundred boosting rounds.

Within each fold, we construct three sets of control variables. First, restaurant baseline helpfulness. For every restaurant that appears in the training data, we compute the average number of helpful votes its reviews receive in the training split. We assign that average to all reviews from the same restaurant in both the training and validation portions of the fold. If a restaurant appears only in the validation portion, we assign the overall training-set average instead. This variable absorbs persistent differences across restaurants (for example, size of audience, location visibility, or platform exposure) that could confound the effect of review text. Second, early-review indicator. We order reviews within each restaurant by date and count how many reviews had been posted for that restaurant before the focal review. We then create a binary flag that equals one when the focal review was written during the restaurant's early period, defined as fewer than five prior reviews, and zero otherwise. This captures systematic differences in helpfulness early in a restaurant's life cycle, when readers have less information and engagement patterns can differ. Third, review-level star-rating dummies. We include four binary indicators for one-, two-, three-, and four-star reviews, leaving five stars as the baseline category. These dummies control for the numeric rating of the review so that the contribution of text features is not confounded with rating valence. We append these controls to the sparse text-feature matrix and use the combined design matrix for model fitting and SHAP attribution.

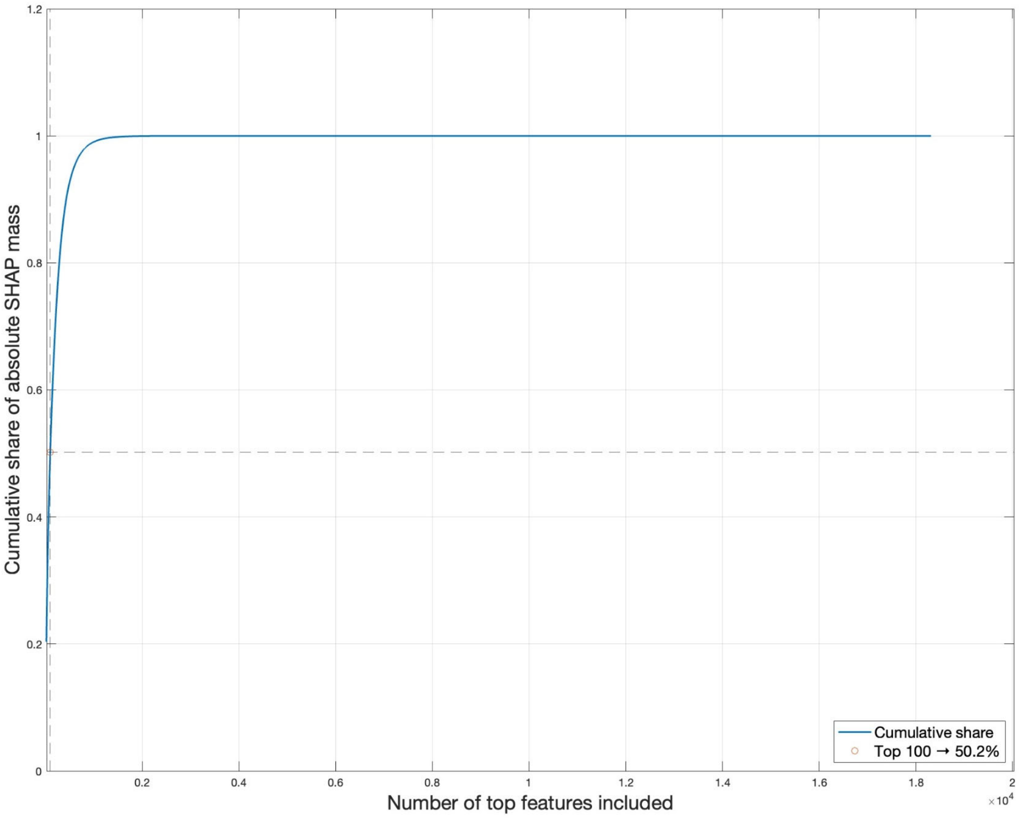
For attribution, we use SHAP's tree explainer with the fold's training design matrix as background and compute SHAP values on held-out validation reviews. We summarize each feature's global importance as the absolute value of its out-of-sample mean SHAP computed across folds, so every review contributes exactly once. This reduces optimism bias because training-set SHAP can inflate importance for features that the model overfits (Cawley and Talbot 2010; Lundberg and Lee 2017). The SHAP importance distribution in our data is relatively diffuse (Figure 2).

Figure 2: Distribution of Feature SHAP Importance with Log-Scaled X-axis.



To maintain interpretability, we restrict our interpretation effort to the top 100 features, ranked by global importance. This subset of features jointly accounts for more than 50% of the total absolute mean SHAP mass (Figure 3). Each feature activates in more than 1% of reviews, and membership in the top 100 is stable across folds. This procedure prioritizes high-impact, well-supported effects and limits over-interpretation of numerous weak or redundant signals.

Figure 3: Cumulative Absolute SHAP Importance by Top-N Features.



Interpreting Features

We applied our recognition procedures to interpret the top-performing features from the previous analysis. The features in this stage spanned a variety of content types. Some features revealed specific rhetorical patterns. For example, one feature activated on punctuation after simple evaluations roughly of the form [subject] is [evaluative adjective] (e.g., “the food is amazing,” “the music is loud,” but not “the menu is seasonal.”). Another captured a specific formula for emphatic complaints (e.g., “Never in my life,” and “What in god’s name”). And yet another captured users acknowledging a reader’s implicit responses with phrases starting with words like “yes” (e.g., “It’s the best restaurant in the city. Yes, it’s expensive. But it’s worth it.”).

Others revealed specific categories of words. For example, one feature activated in the presence of profanity. Another activated in the presence of adjectives that describe food hyperbolically, often using explosion or expansion related metaphors (e.g., “bursting,” “blast,” “loaded”). And yet another activated on words describing a restaurant’s atmosphere (particularly the word ‘atmosphere’ itself, along with related terms like ‘ambiance,’ ‘environment,’ and ‘vibe’).

Some features activated on specific uses of punctuation. One activated on dash-like punctuation (i.e., dashes, double-dashes, em-dashes) used for parenthetical insertions or dramatic breaks (e.g., “Amazing food -- we’ll definitely return,” and “The location - right downtown - is convenient.”). It did not activate when these characters were used as hyphenations.

In this process, we identified two paralinguistic features strongly associated with review helpfulness: formatted lists and structural discourse markers. We examine these features in more depth below.

Formatted lists

In the feature recognition stage, the strongest activations for this feature were on tokens representing line-breaks. Of the 37,272 times the feature activated, 87.8% were on single and double line-breaks. Further, among all tokens, the line-break tokens had the highest average activation of all tokens.

Next, we examined reviews containing the strongest activations of this feature. Among reviews with the strongest activating tokens, all involved formatted lists where list items were separated by line-breaks. In addition, activations tended to occur on all line-breaks except the first in a list (See examples [1] and [2] in Table 1).

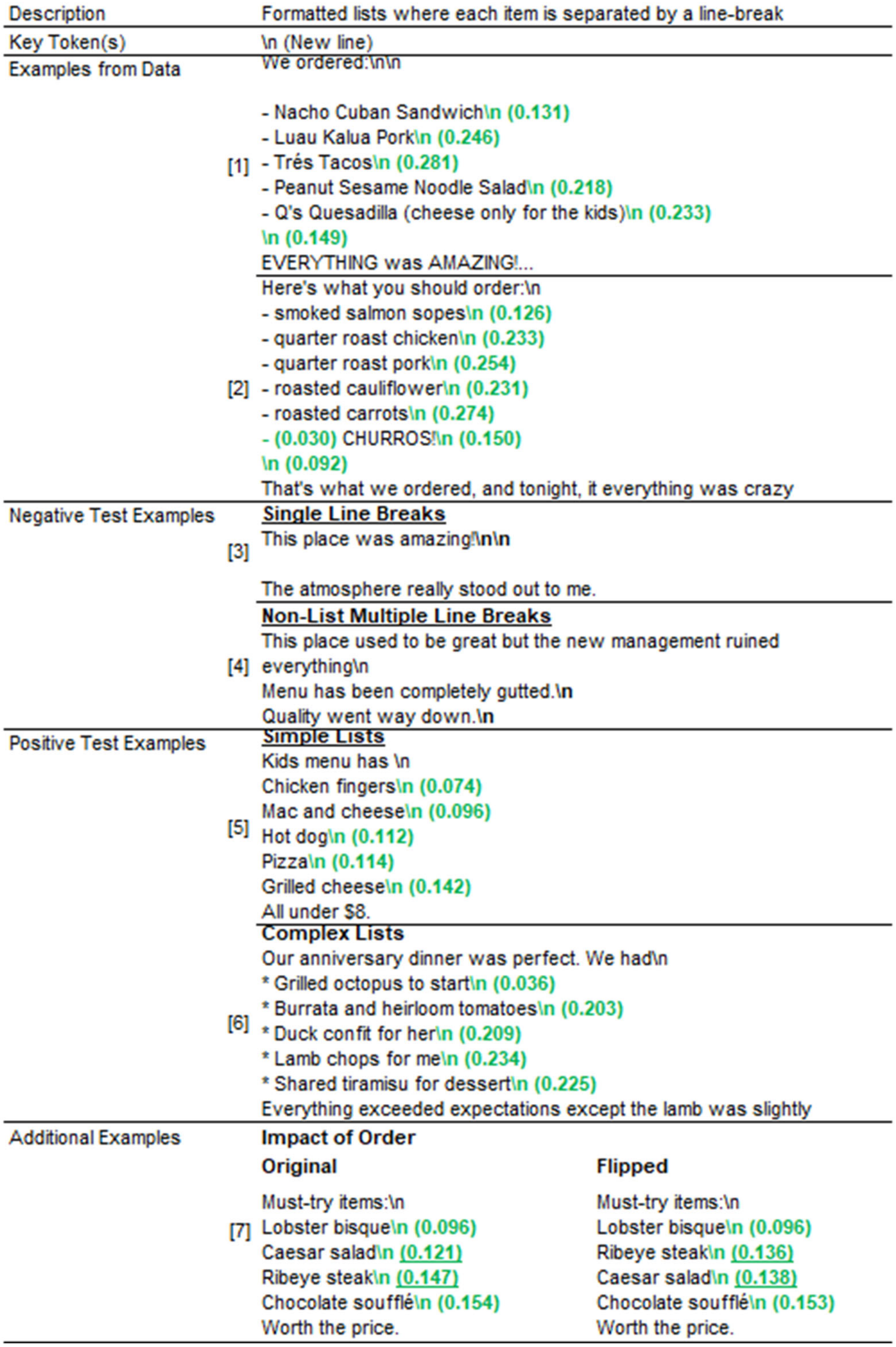
We then proceeded to model-based validation to understand the circumstances under which a new line character activated the feature. We started by examining text that included line-breaks but were not lists. We created 10 reviews that included a single line-break. These line-breaks could be single or multiple lines. As expected, this feature did not activate in any of these cases showing that the feature only activates on repeated line-breaks interspersed between text (See example [3] in Table 1).

We also created 10 reviews that had multiple line-breaks that were not lists. Among these test reviews, 8 had no activations and 2 had very weak activations compared to the activations on lists we saw for lists in the feature-recognition stage (See example [4]). This test suggests that the feature does not activate in response to a review having multiple paragraphs or sentences separated by line-breaks.

Next, we examined text that we expected to activate the feature strongly. We created 10 reviews that contained simple lists of food items and found that in all cases, the feature activated on all line-breaks except for the first one (Example [5]). We also looked at 10 lists that had more complex grammatical constructions and found similar results (Example [6]).

Finally, we generated additional reviews to help us further understand the behavior of the feature. Examining activations on lists, we noted that activation strength increased for later items. To test whether position mattered, we created 5 pairs of reviews where the middle two items in the lists were flipped. In all cases, the activation levels were higher when an item was moved later in the list (Example [7]).

Table 1: Interpretability Examples for Formatted List Feature.



Description		Formatted lists where each item is separated by a line-break	
Key Token(s)		\n (New line)	
Examples from Data	[1]	We ordered:\n\n - Nacho Cuban Sandwich**\n (0.131)** - Luau Kalua Pork**\n (0.246)** - Trés Tacos**\n (0.281)** - Peanut Sesame Noodle Salad**\n (0.218)** - Q's Quesadilla (cheese only for the kids)**\n (0.233)** **\n (0.149)** EVERYTHING was AMAZING!...	
	[2]	Here's what you should order:\n - smoked salmon sopes**\n (0.126)** - quarter roast chicken**\n (0.233)** - quarter roast pork**\n (0.254)** - roasted cauliflower**\n (0.231)** - roasted carrots**\n (0.274)** **- (0.030)** CHURROS!**\n (0.150)** **\n (0.092)** That's what we ordered, and tonight, it everything was crazy	
Negative Test Examples	[3]	**Single Line Breaks** This place was amazing!**\n\n** The atmosphere really stood out to me.	
	[4]	**Non-List Multiple Line Breaks** This place used to be great but the new management ruined everything\n Menu has been completely gutted.**\n** Quality went way down.**\n**	
Positive Test Examples	[5]	**Simple Lists** Kids menu has \n Chicken fingers**\n (0.074)** Mac and cheese**\n (0.096)** Hot dog**\n (0.112)** Pizza**\n (0.114)** Grilled cheese**\n (0.142)** All under $8.	
	[6]	**Complex Lists** Our anniversary dinner was perfect. We had\n * Grilled octopus to start**\n (0.036)** * Burrata and heirloom tomatoes**\n (0.203)** * Duck confit for her**\n (0.209)** * Lamb chops for me**\n (0.234)** * Shared tiramisu for dessert**\n (0.225)** Everything exceeded expectations except the lamb was slightly	
Additional Examples	[7]	**Impact of Order** **Original** Must-try items:\n Lobster bisque**\n (0.096)** Caesar salad**\n (0.121)** Ribeye steak**\n (0.147)** Chocolate soufflé**\n (0.154)** Worth the price.	 **Flipped** Must-try items:\n Lobster bisque**\n (0.096)** Ribeye steak**\n (0.136)** Caesar salad**\n (0.138)** Chocolate soufflé**\n (0.153)** Worth the price.

Note. Bolded green text indicates activated tokens. Numbers in brackets are activation levels.

These analyses suggest that this feature captures the presence of structured, formatted lists within reviews, with a progressive activation pattern that intensifies for later list items. The feature distinguishes between genuine list structures and mere paragraph breaks or isolated line-breaks, suggesting it has learned specifically to identify itemized lists.

Structural discourse markers

The strongest activations for this feature were on tokens indicating that the review had been changed since its initial posting. These included tokens like “Update,” “Edit,” and “ETA” (edited to add). Activations included variations of these tokens in different cases (e.g., “UPDATE” and “Update”) as well as variations of these words (e.g., Edit and Edited). The feature also activated less strongly on structural discourse markers indicating conclusions (e.g., “Overall” and “Conclusion”) and even more weakly on discourse markers for transitions (e.g., “Also” and “Lastly”).

The feature activated 16,541 times. Of the 15,965 (96.5%) activations involving tokens that activated more than 10 times, 53.1% indicated conclusions, 27.2% indicated transitions, and 9.2% indicated updates. The remainder could not be readily identified. When looking at average activations, weighted by frequency, update tokens had the strongest activations (.110), conclusion tokens had weaker activations (.062), and transition tokens had the weakest (.048).

Examining reviews containing the strongest activations of this feature, we found that all involved double line-breaks followed by variations of the token “Update.” This suggests that strong activations involve declaring an update in a clearly separate paragraph of the review. Examples [1] and [2] in Table 2 are examples of reviews where this token activates.

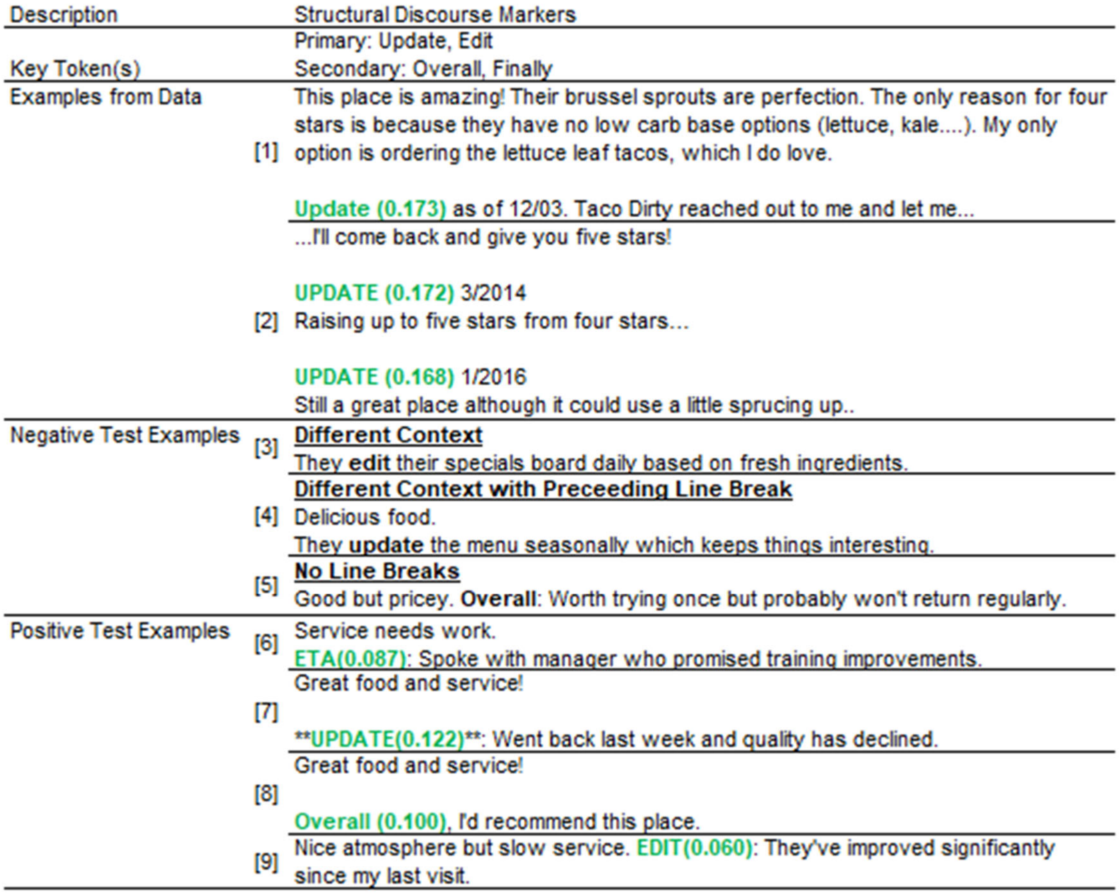
In model-based validation, we first examine cases where strongly activating words were used in different contexts. We created 10 reviews which used strongly activating tokens like "update" and "edit" in different contexts within sentences. As expected, the feature did not activate with any of these reviews (Examples [3] and [4]). Notably, the feature does not activate for the use of "update" soon after a line-break in a different context (Example [4]).

Next, we examine the impact of line-breaks. We created 3 nearly identical sets of 10 reviews. One set used double line-breaks prior to the target token, common in strong activations, one used a single line-break, and one used no line-breaks. We find that the tokens activate after both double and single line-breaks. When we removed line-breaks altogether, weaker activations occurred on tokens that previously activated strongly (e.g., "EDIT"; Example [9]), but not on other tokens (e.g., "Overall"; Example [5]).

To examine reviews where we expected the token to activate strongly, we tested 10 reviews that contained discourse markers that activated the feature in our data separated by one or two line-breaks. In all cases, tokens activated the feature as expected (Example [6], [7], and [8]). All tokens activated as expected, including slight variations (e.g., "**UPDATE**").

Overall, this feature detects structural discourse markers: combinations of words and formatting that signal a transition between sections of a review. It activates most strongly for post-hoc updates and edits, moderately for conclusions, and weakly for transitions. The feature demonstrates contextual sophistication. For example, tokens like "UPDATE" trigger strong activation after line-breaks but fail to activate when appearing within regular sentences. This selectivity shows the feature distinguishes genuine structural discourse from incidental word usage, capturing how reviewers organize their thoughts in text.

Table 2: Interpretability Examples for Structural Discourse Marker Feature.



Description		Structural Discourse Markers
Key Token(s)		Primary: Update, Edit Secondary: Overall, Finally
Examples from Data	[1]	This place is amazing! Their brussel sprouts are perfection. The only reason for four stars is because they have no low carb base options (lettuce, kale....). My only option is ordering the lettuce leaf tacos, which I do love. **Update (0.173)** as of 12/03. Taco Dirty reached out to me and let me...
	[2]	...I'll come back and give you five stars! **UPDATE (0.172)** 3/2014 Raising up to five stars from four stars… **UPDATE (0.168)** 1/2016 Still a great place although it could use a little sprucing up..
Negative Test Examples	[3]	**Different Context** They **edit** their specials board daily based on fresh ingredients.
	[4]	**Different Context with Preceeding Line Break** Delicious food. They **update** the menu seasonally which keeps things interesting.
	[5]	**No Line Breaks** Good but pricey. **Overall**: Worth trying once but probably won't return regularly.
Positive Test Examples	[6]	Service needs work. **ETA(0.087)**: Spoke with manager who promised training improvements.
	[7]	Great food and service! **UPDATE(0.122)**: Went back last week and quality has declined.
	[8]	Great food and service! **Overall (0.100)**, I'd recommend this place.
	[9]	Nice atmosphere but slow service. **EDIT(0.060)**: They've improved significantly since my last visit.

Note. Bolded green text indicates activated tokens. Numbers in brackets are activation levels.

Experimental Validation

Experiment 1: formatted list

Design and procedure. We obtained a sample of 340 participants (mean age = 41.5, std = 13.8; 53.5% female) from Prolific after removing the responses of participants who failed the attention check. Participants received a flat rate of $1.40 for completing the study. The study was pre-registered at AsPredicted #244033 (https://aspredicted.org/bjqg-23hv.pdf).

Participants were randomly assigned to view one of four restaurant review pairs. Each pair contained two versions of the same review: one written in a structured format using a formatted list and the other written in continuous prose. The two reviews were presented side-by-side, as shown in Figure 4. As the dependent measure, participants were asked to imagine they were browsing reviews for this restaurant and make a binary choice on which of these two reviews they are more likely to find helpful.

Result. Across all participants, 60.8% selected the review with a formatted list as more helpful. A one-sample proportions test showed that the percentage of structured reviews being chosen is significantly higher than 50% (χ^2 = 15.6, df = 1, p < .001).

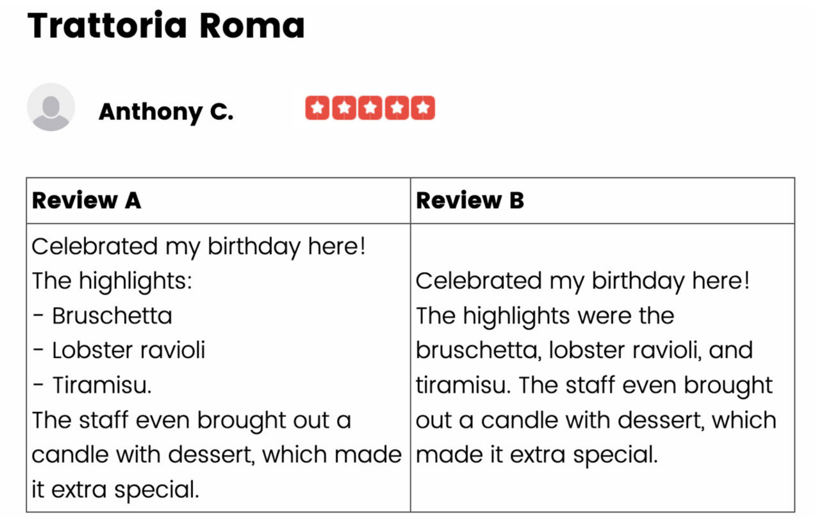
Experiment 2: structural discourse marker

Design and procedure. We obtained a sample of 365 participants (mean age = 43.0, std = 12.9; 57.5% female) from Prolific after removing the responses of participants who failed the attention check. Participants received a flat rate of $1.40 for completing the study. The study was pre-registered at AsPredicted #244087 (https://aspredicted.org/mdqn-t7w9.pdf).

Following the same procedure as in Experiment 1, participants were randomly assigned to view one of four restaurant review pairs. Each pair contained two versions of the same review: one with a line-break and a sentence-initial discourse marker at the start of the following paragraph (e.g., “UPDATE”, “Overall”, etc.) and one without such format. The two reviews were presented side-by-side. Participants were asked to imagine they are browsing reviews for this restaurant and make a binary choice on which of these two reviews they are more likely to find helpful.

Result. 71.8% participants selected the review with a discourse marker as more helpful. A one-sample proportions test against a null hypothesis showed that the percentage of structured reviews being chosen is significantly higher than 50% (χ^2 = 68.4, df = 1, p < .001).

Figure 4: Example of Stimuli for Experimental Studies.



Trattoria Roma

Anthony C. ★★★★★

Review A	Review B
Celebrated my birthday here! The highlights: - Bruschetta - Lobster ravioli - Tiramisu. The staff even brought out a candle with dessert, which made it extra special.	Celebrated my birthday here! The highlights were the bruschetta, lobster ravioli, and tiramisu. The staff even brought out a candle with dessert, which made it extra special.

Comparison With Other Methods

Next, we compared our results with those from two commonly used topic-modeling methods: LDA and clustered embeddings.

LDA

We compared our method with a standard Latent Dirichlet Allocation (LDA) approach, in line with prior work that builds brand or attribute maps from reviews using unsupervised topics (Tirunillai and Tellis 2014). LDA assumes a bag-of-words representation in which word order and dependencies are ignored. We use it here as a benchmark to assess whether and how its insights differ from those of our method.

We implemented LDA on the full dataset, varied the number of topics from 2 to 10 and evaluated models using coherence (Röder et al. 2015). Coherence is the average cosine similarity between Normalized Pointwise Mutual Information (NPMI)–based sliding-window co-occurrence vectors for a topic's top words. It serves as a proxy for human interpretability, rewarding topics whose top words co-occur more than chance and penalizing collections of unrelated words. Coherence peaked at nine topics, which we report as our primary LDA specification (Table 3).

Table 3: Topics Extracted from Our Review Corpus Using LDA.

Topic	Representative phrases
1	bad; bland; dry; cold; charge; plate; cook; tell
2	taco; tacos; fish; al pastor; shrimp; Tuesday; favorite; delicious
3	service; margarita; staff; friendly; atmosphere; excellent; highly recommend; drink
4	breakfast; tamale; truck; Tucson; lunch; egg; stop; favorite
5	salsa; enchilada; fajita; chip salsa; authentic mexican; price; lunch; fresh
6	sauce; flavor; spicy; guacamole; shrimp; pork; cheese; dish
7	bar; beer; drink; happy hour; music; brunch; cocktail; atmosphere
8	table; wait; minute; manager; rude; server; seat; customer
9	burrito; Chipotle; bowl; Qdoba; Taco Bell; rice; bean; veggie

The LDA model produces intuitive but broad themes and yields a useful strategic map of issues (e.g., tacos vs. burritos; bar/ambience; wait/service incidents). However, because it relies on bag-of-words co-occurrence, it is less effective at uncovering finer language patterns and paralinguistic cues that matter for how readers evaluate reviews. In our context, LDA could not reliably isolate cues such as punctuation, formatted lists, and discourse markers. In contrast, our method begins with contextual embeddings that capture semantics beyond surface form and then learns sparse, disentangled features that map cleanly onto nuanced communicative features. In practice, these features are sharper, less overlapping, and more useful for downstream analyses, whereas LDA provides a high-level map but not the linguistic resolution our application requires.

BERTopic

We also compared our method with BERTopic (Grootendorst 2022). BERTopic is a neural topic-modeling workflow widely used to discover coherent, human-readable themes in large text collections. The method embeds each review with a sentence-transformer, reduces dimensionality with UMAP (Uniform Manifold Approximation and Projection), clusters embeddings with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), and represents each cluster with class-based TF-IDF over an n-gram vectorizer so labels include meaningful phrases. Unlike LDA, BERTopic does not require a preset number of topics: HDBSCAN infers them by extracting stable density-based clusters and labeling sparse points as noise (Campello et al. 2013). Applied to our review dataset, BERTopic produced 174 topics (excluding outliers).

Compared to LDA, the resulting topics are relatively fine-grained, contextually sensitive, and intuitive, producing clusters like beverages, occasions, dishes, operational issues, chains and brands. Table 4 reports the ten highest-volume topics and their representative phrases. It reliably separates dishes, occasions, brand clusters, and operational issues, which LDA tended to blend. However, BERTopic's labels do not explicitly capture paralinguistic cues and micro-language patterns that shape reader judgments. Our method complements and goes beyond BERTopic by learning a sparse, multi-feature representation per review that disentangles features such as punctuation and emphasis, formatted lists, discourse markers, and conversational interjections like “oh my”. Moreover, our method allows multiple co-occurring features within one text, unlike BERTopic’s default single-label clustering[1].

Table 4: Ten Highest-Volume Topics Using BERTopic.

Topic	Representative phrases
1	margaritas; margarita; tequila; drinks; service; food
2	tacos; taco; best tacos; delicious; love; amazing
3	burrito(s); breakfast burrito(s); beans; rice; meat
4	mexican food; authentic; authentic mexican; restaurant
5	chipotle; burrito bowl; line; location; order
6	table; minutes; asked; told; server; manager; hostess
7	tucson; mexican food tucson; salsa
8	nashville; east nashville; tacos
9	beach; burger; crab; shrimp; grouper
10	taco bell; drive; worst taco bell; fast food

Conclusion

In this article, we developed a methodological framework to uncover highly contextual features from reviews that impact their perceived helpfulness. By looking at many fine-grained features in parallel, we have identified concrete and actionable features that would be difficult to recover with traditional natural language processing (NLP) models. In particular, we found two highly influential textual paralinguistic cues which would have been difficult to detect using

[1] BERTopic can output per-topic probabilities when enabled, and one can threshold these to create multi-label assignments. However, this requires an arbitrary cutoff and does not produce the disentangled, monosemantic features our method recovers.

other methods. In addition to identifying these features in our dataset of 432,248 reviews, we also validated their impact on review helpfulness experimentally.

Theoretical and Practical Implications

Despite playing a key part in online commerce, relatively little is understood about textual drives of engagement with reviews because they are unstructured. A key contribution of our research is to develop a method that can detect highly influential, but heretofore hidden drivers of engagement with these reviews.

Our findings make two contributions. First, we apply MI methods to marketing text analytics, enabling scalable detection of thousands of semantic, lexical, and structural features and estimating their relative contributions to outcomes in large text databases. Unlike traditional NLP methods that examine only a few features at a time (requiring meta-analysis for comparison) or topic modeling (which operates at a different level of abstraction), our bottom-up approach captures context-dependent patterns including typically ignored elements like spacing, line-breaks, and textual paralanguage. By examining thousands of candidates in parallel while controlling for other variables, this process identifies not just impactful features but the most impactful ones. Further by validating selected high-importance features through both models and experiments, we confirm their importance. The fine-grained features we detect and validate provide concrete recommendations for practitioners (e.g., specific formatting choices that drive engagement) while complementing existing methods. For instance, our features can be mapped to topics from topic modeling to reveal their defining microfeatures. As our paralinguistics analysis demonstrates, this approach uncovers engagement drivers that traditional methods miss, offering both theoretical advances and practical applications.

Second, we detect concrete textual paralinguistic cues which are difficult to detect with existing methods. In particular, we find that structured lists, where each item is separated by a line-break, and structural discourse markers, which are combinations of words and formatting that signal a transition between sections of a review (e.g., “UPDATE” and “Conclusion”) positively impact helpfulness. Notably, these features are much more fine-grained than those identified by existing topic modeling methods like LDA or clustered embeddings.

The concrete and fine-grained nature of the features we uncover through our framework is of practical importance because it allows us to provide platforms, businesses on platforms, and content generators with specific and actionable advice for understanding, and crucially, generating engaging content. When appropriate, any user or brand interested in creating helpful reviews should use formatted lists that separate items with line-breaks. Further, they should structure their reviews clearly by combining line-breaks with specific words that signal to the reader that they are entering a new section of the review. Further, including clearly marked sections for updates and edits has a particularly strong influence on helpfulness. As platforms increasingly rely on generative AI to generate content like summaries, they, too, can benefit from this guidance. Finally, platforms can use these features to identify and surface reviews that consumers are likely to find helpful.

While we focused on detecting textual paralinguistics and helpfulness in restaurant reviews, our approach can be applied to many contexts with rich textual databases. We focused on textual paralinguistics because they are hard to detect with other methods, but this method also identifies other classes of features including lexical constructions, concepts, and word types. During feature interpretation, we discovered features detecting profanity, phrases previewing emphatic negatives (e.g., “never in my life,” “why in the world”), directive statements with “your” (e.g., “Save your money” but not “They will save your money”), and words related to the environment of the restaurant (e.g., “environment” and “vibe”). This flexibility enables exploration of diverse questions like how textual features vary across product categories, what features define reviews of differing valence, or what predicts social media engagement like likes and shares. While we focused on review helpfulness, these methods could readily examine other outcomes across platforms. Further, the methods could be applied more broadly to analyze other textual data like customer service transcripts, corporate communications, or earnings call transcripts.

It is also worth noting that the methods we employed in this article will continue to improve as technology evolves. While we applied our method to a specific LLM, in principle these methods can be applied to future LLMs with more capabilities. In fact, as LLMs grow more capable, it is likely that they will detect a richer array of features (e.g., compare Bricken et al. 2023 to Templeton et al. 2024) from a wider array of modalities (i.e., images, audio, video). And, as computational capacity grows, these methods will become more accessible to both researchers and practitioners.

Limitations and Future Research

Our research has several limitations which point to opportunities for future research. First, while we focused on textual paralinguistic features because they were difficult for other methods to detect, there are likely other drivers of review helpfulness uncovered by our framework that researchers are unaware of. Future research could examine the impact of these additional features on engagement using the SAE we created for our research.

Second, our approach requires manual interpretation of features by humans. While manual interpretation is commonly employed in topic modeling, it potentially has two problems.

First, human judgment can introduce bias. We limited the impact of this bias through model-based and experimental validations. Second, manual interpretation and the validation that accompanies it are slow. Because this method generates thousands of potentially useful features, automated interpretation will be necessary to fully exploit its capabilities. There are some efforts underway to automate this type of feature analysis using LLMs (e.g., Paulo et al. 2024).

Third, we leverage token representations from one layer of GPT2, which is a relatively small LLM. We used GPT2 because it is commonly used in MI. However, future research could use more advanced open-source models (e.g., Llama 4 or GPT-oss) to generate even richer sets of features, particularly as the cost of computation and data storage declines. Future research could examine activations from multiple sections of an LLM simultaneously. Since LLM representations grow increasingly abstract in later layers, examining multiple layers would reveal feature hierarchies from concrete to abstract. Alternatively, training SAEs with different hidden layer sizes on the same activations could generate hierarchies, as larger hidden layers often split features into more specific subcategories (Templeton et al. 2024).

Finally, future research could use larger datasets to train SAEs. Our dataset contained approximately 52 million tokens, sufficient to produce fine-grained and interpretable features. However, expanded training data would likely yield more conceptually sophisticated features. For example, Templeton et al. (2024) found that models trained on substantially larger multimodal datasets developed features activating for the Golden Gate Bridge (across both text and images), transit infrastructure, and tourist attractions. Larger training datasets may enable detection of such conceptually rich features.

References

Abdurahman, Suhaib, Alireza S. Ziabari, Alexander K. Moore, Daniel M. Bartels and Morteza Dehghani (2025), A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, *8*(2), 25152459251325174.

Ahn, Jisoo, Hyunsang Son, and Arnold Dongwoo Chung (2021), "Understanding public engagement on twitter using topic modeling: The 2019 Ridgecrest earthquake case," *International Journal of Information Management Data Insights*, 1 (2), 100033.

Anderson, Michael and Jeremy Magruder (2012), "Learning from the crowd: Regression discontinuity estimates of the effects of an online review database: Learning from the crowd," *Economic Journal*, 122 (563), 957–89.

Arora, Neeraj, Ishita Chakraborty, and Yohei Nishimura (2025), "AI–human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators," *Journal of Marketing*, 89 (2), 43–70.

Ba, Jimmy. L., Jamie R. Kiros, & Geoffrey E. Hinton (2016), Layer normalization. *arXiv preprint arXiv:1607.06450*.

Babić Rosario, Ana, Francesca Sotgiu, Kristine De Valck, and Tammo H. A. Bijmolt (2016), "The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors," *Journal of Marketing Research*, 53 (3), 297–318.

Bai, Hui, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer (2025), "LLM-generated messages can persuade humans on policy issues," *Nature Communications*, 16 (1), 6037.

Baskin, Kara (2022), "Burgers with bugs? What happens when restaurants ignore online reviews," *Harvard Business School Working Knowledge*, https://www.library.hbs.edu/working-knowledge/what-happens-when-restaurants-ignore-online-reviews.

Bengio, Yoshua, Ian Goodfellow and Aaron Courville (2016), *Deep learning*. Cambridge, MA, USA: MIT press.

Berger, Jonah and Katherine L. Milkman (2012), "What makes online content viral?," *Journal of Marketing Research*, 49 (2), 192–205.

Bickart, Barbara, & Robert M. Schindler (2001). Internet forums as influential sources of consumer information. *Journal of interactive marketing*, 15(3), 31-40.

Brandwatch (2024), "Brandwatch advances brand reputation management with innovative proprietary and generative AI integration," *Brandwatch*. https://www.brandwatch.com/press/press-releases/brandwatch-advances-brand-reputation-management-with-innovative-proprietary-and-generative-ai-integration.

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah (2023), Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Anthropic*. https://transformer-circuits.pub/2023/monosemantic-features/index.html

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini

Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020), “Language models are few-shot learners,” in *Proceedings of the Advances in Neural Information Processing Systems*, 33, 1877–901.

Burnap, Alex, John R. Hauser, and Artem Timoshenko (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science*, 42(6), 1029-1056.

Campello, Ricardo JGB, Davoud Moulavi, and Jörg Sander (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160-172.

Cao, Qing, Wenjing Duan, and Qiwei Gan (2011), “Exploring determinants of voting for the ‘helpfulness’ of online user reviews: A text mining approach,” *Decision Support Systems*, 50 (2), 511–21.

Cawley, Gavin C., and Nicola LC Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 11: 2079-2107.

Ceylan, Gizem, Kristin Diehl, and Davide Proserpio (2024) "Words meet photos: When and why photos increase review helpfulness." *Journal of Marketing Research* 61(1), 5-26.

Challa, Uma (2025), “Customer service AI: Hone in on high-ROI use cases,” *Gartner*. https://www.gartner.com/en/articles/customer-service-ai.

Chen, Pei-Yu, Samita Dhanasobhon, and Michael D. Smith (2008), “All reviews are not created equal: The disaggregate impact of reviews and reviewers at Amazon.com,” *SSRN*, http://dx.doi.org/10.2139/ssrn.918083.

Chen, Tianqi and Carlos Guestrin (2016), XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794.

Christy, Cheung and Lee Matthew (2012), “What drives consumers to spread electronic word of mouth in online consumer-opinion platforms,” *Decision Support Systems*, 53 (1), 218–25.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey (2024), *Sparse autoencoders find highly interpretable features in language models.* In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Dagogo-Jack, Sokiente W., and Jared J. Watson (2025), “Most Read Versus Most Shared: How Less (vs. More) Social Popularity Labels Influence News Media Consumption,” *Journal of Consumer Research*, advance article

Deniz, Begum Karaci, Chandra Gnanasambandam, Martin Harrysson, Alharith Hussin, and Shivam Srivastava (2023), “Unleashing Developer Productivity with Generative AI,” *McKinsey & Company,* (June 27), https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Dew, Ryan, Asim Ansari, & Olivier Toubia (2022). Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science*, 41(2), 401-425.

Ethayarajh, Kawin (2019), How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Fang, Limin (2022), “The effects of online review platforms on restaurant revenue, consumer learning, and welfare,” *Management Science*, 68 (11), 8116–43.

Farace, Stefania, Francisco Villarroel Ordenes, Dennis Herhausen, Dhruv Grewal, and Ko de Ruyter (2025), “Standing out while fitting in: Visual design of text overlays in social media communication,” *Journal of Marketing*, 00222429251322773.

Gao, Leo, Tom D. la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu (2024), Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Ghose, Anindya and Panagiotis G. Ipeirotis (2010), Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE transactions on knowledge and data engineering*, *23*(10), 1498-1512.

Goli, Ali and Amandeep Singh (2024), “Frontiers: Can large language models capture human preferences?,” *Marketing Science*, 43 (4), 709–22.

Goldsmith, Ronald E., and David Horowitz (2006). Measuring motivations for online opinion seeking. *Journal of interactive advertising*, 6(2), 2-14.

Grootendorst, Maarten (2022), BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Harkar, Shalini (2025), “What is Vibe Coding?,” *IBM*, https://www.ibm.com/think/topics/vibe-coding.

Harkness, Lisa, Kelsey Robinson, Eli Stein, and Winnie Wu (2023), “How generative AI can boost consumer marketing,” *Mckinsey*, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/how-generative-ai-can-boost-consumer-marketing.

Hartmann, Jochen, Juliana Huppertz, Christina Schamp, and Mark Heitmann (2019) "Comparing automated text classification methods." *International Journal of Research in Marketing* 36(1), 20-38.

Herhausen, Dennis, Stephan Ludwig, Dhruv Grewal, Jochen Wulf, and Marcus Schoegel (2019), “Detecting, Preventing, and Mitigating Online Firestorms in Brand Communities,” *Journal of Marketing*, 83 (3), 1–21.

Huang, Peng, Nicholas H. Lurie, and Sabyasachi Mitra (2009), “Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods,” *Journal of Marketing*, 73 (2), 55–69.

Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman (2023), “Human heuristics for AI-generated language are flawed,” *Proceedings of the National Academy of Sciences*, 120 (11), e2208839120.

Kanuri, Vamsi K., Christian Hughes, and Brady T. Hodges (2024), “Standing out from the crowd: When and why color complexity in social media images increases user engagement,” *International Journal of Research in Marketing*, 41 (2), 174–93.

Kaul, Rupali, Stephen J. Anderson, Pradeep K. Chintagunta, and Naufel Vilcassim (2025). Call me maybe: does customer feedback seeking impact nonsolicited customers?. *Marketing Science*, 44(1), 129-154.

Kim, Soo-Min Patrick Pantel, Tim Chklovski, and Marco Marco Pennachiotti (2006), “Automatically assessing review helpfulness,” in *Proceedings of the 2006 Conference on empirical methods in natural language processing*, 423–30.

King, Robert Allen, Pradeep Racherla, and Victoria D. Bush (2014) "What we know and don't know about online word-of-mouth: A review and synthesis of the literature." *Journal of interactive marketing* 28(3), 167-183.

Kreps, Sarah, R. Miles McCain, and Miles Brundage (2022), “All the news that’s fit to fabricate: AI-generated text as a tool of media misinformation,” *Journal of Experimental Political Science*, 9 (1), 104–17.

Kristopher, Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, Traci Freling (2014), “How online product reviews affect retail sales: A meta-analysis,” *Journal of Retailing*, 90 (2), 217–32.

Krugmann, Jan Ole and Jochen Hartmann (2024), “Sentiment analysis in the age of Generative AI,” *Customer Needs and Solutions*, 11 (1), 3.

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton (2015), Deep learning. *Nature*, *521*(7553), 436-444.

Lee, Sang G., Silvana Trimi, and Chang-Gyu Yang (2018), Perceived usefulness factors of online reviews: a study of Amazon.com. *Journal of computer information systems*, *58*(4), 344-352.

Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary (2024), “Frontiers: Determining the validity of Large Language Models for automated perceptual analysis,” *Marketing Science*, 43 (2), 254–66.

Li, Yiyi and Ying Xie (2020), “Is a picture worth a thousand words? An empirical study of image content and social media engagement,” *Journal of Marketing Research*, 57 (1), 1–19.

Loshchilov, Ilya, and Frank Hutter (2017), "Decoupled weight decay regularization." *arXiv preprint arXiv:1711.05101*.

Luangrath, Andrea W., Joann Peck, and Victor A. Barger (2017), Textual paralanguage and its implications for marketing communications. *Journal of Consumer Psychology*, *27*(1), 98-107.

Luangrath, Andrea W., Yixiang Xu, and Tong Wang (2023), Paralanguage classifier (PARA): An algorithm for automatic coding of paralinguistic nonverbal parts of speech in text. *Journal of Marketing Research*, *60*(2), 388-408.

Luca, Michael (2016), “Reviews, reputation, and revenue: The case of yelp.com,” SSRN, https://dx.doi.org/10.2139/ssrn.1928601.

Luca, Michael, Abhishek Nagaraj, and Gauri Subramani (2023), “Getting on the map: The impact of online listings on business performance,” *National Bureau of Economic Research*, No. w30810.

Lundberg, Scott and Su-In Lee (2017), A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Ma, Liye and Baohong Sun (2020), Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3), 481–504.

Motoyama, Yasuyuki, and Kareem Usher (2020), “Restaurant Reviews and Neighborhood Effects,” *Papers in Applied Geography*, 6 (4), 386–401.

Mudambi and Schuff (2010), “Research note: What makes a helpful online review? A study of customer reviews on Amazon.com,” *MIS Quarterly*, 34 (1), 185–200.

Mahdikhani, Maryam (2022), “Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic,” *International Journal of Information Management Data Insights*, 2 (1), 100053.

Makhzani, Alireza and Brendan Frey (2013), K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Mancisidor, Rogelio, A., Michael Kampffmeyer, Kjersti Aas, & Robert Jenssen (2021). Learning latent representations of bank customers with the variational autoencoder. *Expert Systems with Applications*, 164, 114020.

Matter, Daniel, Miriam Schirmer, Nir Grinberg, and Jurgen Pfeffer (2024), Investigating the increase of violent speech in incel communities with human-guided gpt-4 prompt iteration. *Frontiers in Social Psychology*, *2*, 1383152.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov (2022), Locating and editing factual associations in gpt. *Advances in neural information processing systems*, *35*, 17359-17372.

Mudambi, Susan M., & David Schuff (2010), Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly*, 185-200.

Nguyen, Peter, Xin (shane) Wang, Xi Li, and June Cotte (2021), “Reviewing experts’ restraint from extremes and its impact on service providers,” *Journal of Consumer Research*, 47 (5), 654–74.

Packard, Grant and Jonah Berger (2017), “How Language Shapes Word of Mouth’s Impact,” *Journal of Marketing Research*, 54 (4), 572–588.

Paulo, Gonçalo, Alex Mallen, Caden Juang, and Nora Belrose (2024) "Automatically interpreting millions of features in large language models." *arXiv* preprint arXiv:2410.13928

Pennebaker, James W., Martha E. Francis, & Roger J. Booth (2001), Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Proserpio, Davide and Giorgos Zervas (2018), “Study: Replying to customer reviews results in better ratings,” *Harvard Business Review.*

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019), Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021), “Learning transferable visual models from natural language supervision,” In *Proceedings of the International Conference on Machine Learning*, 8748–63.

Rajamanoharan, Senthooran, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramár, Rohin Shah, and Neel Nanda (2024), Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*.

Ringel, Daniel (2023), “Creating synthetic experts with generative artificial intelligence,” *SSRN*, https://dx.doi.org/10.2139/ssrn.4542949.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015), Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*: 399-408.

SAS (2024), “GenAI report,” *SAS*, https://www.sas.com/en/offers/dive-shallow.html.

Sedhain, Suvash, Aditya Krishna Menon, Scott Sanner, and Lexing Xie (2015) "Autorec: Autoencoders meet collaborative filtering." In *Proceedings of the 24th international conference on World Wide Web*, pp. 111-112.

Sennrich, Rico, Barry Haddow, & Alexandra Birch (2015), Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Singla, Alex, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall (2024), "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," *Mckinsey*, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024.

Suh, Bongwon, Lichan Hong, Peter Pirolli, and Ed H. Chi (2010), "Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network," in 2010 IEEE Second International Conference on Social Computing, *IEEE*. 177–84

Templeton, Adly, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan (2024), Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Anthropic*. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Timoshenko, Artem, Chengfeng Mao, and John R. Hauser (2025), "Can large language models extract customer needs as well as professional analysts?," *SSRN*.

Timoshenko, Artem, & John R. Hauser (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1), 1-20.

Tirunillai, Seshadri., and Gerard J. Tellis (2014), Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463-479.

Toubia, Olivier, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen (2025) "Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions." *Marketing Science*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017), Attention is all you need. *Advances in neural information processing systems*, *30*.

Villarroel Ordenes, Francisco, Dhruv Grewal, Stephan Ludwig, Ko De Ruyter, Dominik Mahr, and Martin Wetzels (2019), "Cutting through Content Clutter: How Speech and Image Acts Drive Consumer Sharing of Social Media Brand Messages," *Journal of Consumer Research*, 45 (5), 988–1012.

von Zahn, Moritz, Kevin Bauer, Cristina Mihale-Wilson, Johanna Jagow, Maximilian Speicher, and Oliver Hinz (2025), Smart green nudging: Reducing product returns through digital footprints and causal machine learning. *Marketing Science,* 44(4):954-969.

Wang, Kevin, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt (2022), Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Wang, Xin (Shane), Shijie Lu, X. I. Li, Mansur Khamitov, and Neil Bendle (2021), "Audio mining: The role of vocal tone in persuasion," *Journal of Consumer Research*, 48 (2), 189–211.

Wang, Yani, Jun Wang, and Tang Yao (2019), “What makes a helpful online review? A meta-analysis of review characteristics,” *Electronic Commerce Research*, 19 (2), 257–84.

Wu, Chunhua, Hai Che, Tat Y. Chan, and Xianghua Lu (2015), “The economic value of online reviews,” *Marketing Science*, 34 (5), 739–54.

Yazdani, Elham, Anindita Chakravarty, and Jeffrey Inman (2025), “Racial Inequity in Donation-Based Crowdfunding Platforms: The Role of Facial Emotional Expressiveness,” *Journal of Marketing*.

Yin, Dezhi, Samuel D. Bond & Han Zhang (2017). Keep your cool or let it out: Nonlinear effects of expressed arousal on perceptions of consumer reviews. *Journal of Marketing Research*, 54(3), 447-463.

Yoganarasimhan, Hema and Irina Iakovetskaia (2024), “From feeds to inboxes: A comparative study of polarization in Facebook and email news sharing,” *Management Science*, 70 (9), 6461–72.

You, Ya, Gautham G. Vadakkepatt, and Amit M. Joshi (2015), “A meta-analysis of electronic word-of-mouth elasticity,” *Journal of Marketing*, 79 (2), 19–39.

Zhang, Mengxia and Lan Luo (2023), “Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp,” *Management Science*, 69 (1), 25–50.

Zhang, Y. and R. Gosline (2023), “Human favoritism, not AI aversion: People’s perceptions (and bias) toward generative ai, human experts, and human-GAI collaboration in persuasive content generation,” *Judgment and Decision Making*, 18, e41.

Zhu, Yongmin, Miaomiao Liu, Xiaohua Zeng, and Pei Huang (2020), “The effects of prior reviews on perceived review helpfulness: A configuration perspective,” *Journal of Business Research*, 110, 484–94.