Creating Synthetic Specialists with Generative AI

Daniel M. Ringel

Working Paper

October 3, 2024

Abstract

Understanding marketing mix (MMX) effectiveness from consumers' perspectives is critical for marketers as it drives strategic marketing decisions. Identifying which MMX element consumer feedback and online discussion pertain to requires costly human expertise due to the complexity of the MMX construct. We propose a new approach to solving this problem by creating synthetic specialists—task-specific approximations of generative AI models. We show that generative AI can effectively identify MMX elements in consumer-generated content, but its high cost and resource demands prohibit its use at scale. We evaluate performance through agreement with human experts and find that a synthetic specialist outperforms 27 generative AI models on multiple fronts: it achieves 4% higher human expert agreement, is 297 times faster, and has a significantly smaller CO_2 footprint. Additionally, the synthetic specialist is non-proprietary, runs locally, has perfectly replicable outputs, and preserves confidentiality, offering greater control over AI processes and data security. We further show that MMX classification provides richer insights than commonly used sentiment analysis. Studying over 600,000 tweets about 699 brands, we find that sentiment is distributed differently across MMX elements. Deeper analysis identifies different strengths and weaknesses among competitors and reveals consumer pain points in brands' marketing mix.

Keywords: *Generative Artificial Intelligence, Classification, Marketing Mix*

INTRODUCTION

Generative artificial intelligence (AI) has emerged as a transformative technology, focusing on creating new content—such as text, images, or music—by learning patterns from existing data (Radford et al. 2018). Utilizing advanced models like neural networks and algorithms such as generative adversarial networks and transformers, generative AI produces content that often matches human creativity (Van Dis et al. 2023). Its novelty lies in generating original material rather than merely analyzing data, opening new avenues in creativity, automation, and innovation.

In marketing, the potential of generative AI has begun to capture attention. Recent studies demonstrate its use as a substitute for human subjects in market research. For instance, Li et al. (2024) queried GPT-4 for consumer perceptions of automotive brands to generate perceptual maps. Brand, Israeli, and Ngwe (2023) explored the benefits of GPT-3 for understanding consumer preferences, while Castelo et al. (2024) examined GPT 4's creative capabilities in generating innovative product ideas, finding that it can produce ideas rated as more creative than those generated by laypeople and professionals. Additionally, Jürgensmeier and Skiera (2024) used generative AI to provide automated feedback on complex exercises requiring coding, statistics, and economic reasoning.

Despite its promise, generative AI faces significant challenges that hinder its widespread adoption in specialized tasks. The cost of training large AI models has grown exponentially, making it difficult for firms to sustain such investments (Maslej et al. 2023). The computational and energy resources required for inference raise concerns about carbon emissions, with the AI industry becoming a significant contributor to climate change if current trends continue (Sundberg 2024). Privacy risks associated with aggregating datasets necessitate governance to protect sensitive information (Baquero et al. 2020). Moreover, reliance on proprietary models like OpenAI's GPT-4 creates dependence on providers who control access, pricing, and capabilities, thus exposing organizations to confidentiality and data privacy risks (Busch 2023; Daniels 2023). The non-deterministic nature of generative AI responses also hampers reproducibility, a major concern in the research community (Van Noorden and Perkel 2023; Hou and Ji 2024).

We contend that utilizing general-purpose generative AI models for specific tasks such as text classification is inefficient due to their massive resource demands and inherent limitations. General AI models are applicable to many tasks and promise to revolutionize knowledge work, but they are unwieldy, resource-intensive, and largely closed and proprietary (Leffer 2023). In contrast, specialized AI models, such as classifiers built for individual tasks, are known for their efficiency and accuracy, although they require domain expertise for training. This dichotomy raises the question: why employ a vast generalist model when a specialized one suffices?

Organizations increasingly depend on machine learning to extract intelligence from vast amounts of unstructured data, including news and social media, customer interactions, reports, policies, and internal communications (Chakraborty, Kim, and Sudhir 2022). Classification models have garnered significant attention due to their versatility and potency. They swiftly analyze large volumes of data to identify patterns and categorize them into predefined classes, helping firms harness the potential of their unstructured information assets (Frankel, Jennings, and Lee 2022). By identifying constructs of interest such as specific topics, bias and sentiment, compliance, or emotions, classification models unlock hidden insights that managers can leverage to drive business growth (Abbasi et al. 2019; Şeref et al. 2023). They are instrumental in decision-making processes, customer understanding, and risk mitigation across sectors and functions.

The effectiveness of modern classification models relies heavily on their training, which hinges on the availability of a substantial amount of correctly labeled examples (Hartmann et al. 2023). For simple constructs like sentiment, spam, or emotions, analysts can resort to crowdsourced labels (Snow et al. 2008). However, more complex constructs—those with higher levels of abstraction, ambiguity, and multifaceted dimensions—require expert labelers with domain knowledge and specialized skills (Hartmann et al. 2023). Experts are typically a scarce and expensive resource, introducing a substantial bottleneck in the labeling process. This poses a key problem in deploying classification models in marketing to identify fundamental constructs such as the marketing mix, drivers of brand equity, or dimensions of service quality in unstructured data.

This study examines marketing mix elements that consumers discuss online. The marketing mix, or "four Ps of marketing"—product, price, place, and promotion—is at the heart of marketing strategy and is one of the most powerful concepts developed for executives (Kotler et al. 2012; Shapiro 1985). Understanding which element of a brand's marketing mix consumers talk about on social media is important. It guides managers' assessments of their marketing strategy, alerts them to potential risks and opportunities, and identifies which marketing mix levers require attention. However, classifying consumer-generated content regarding marketing mix elements is a complex undertaking. The lack of mutual exclusiveness among the four classes creates ambiguity that labelers must address. While refinements have improved mutual exclusiveness, they often complicate faithful identification by crowdsourced amateurs (Van Waterschoot and Van den Bulte 1992).

To overcome these limitations, we propose using generative AI as a proficient substitute for expert labelers by leveraging its vast body of latent information. Generative AI models are trained on extensive data, including books, reports, news articles, and websites, encompassing many theoretically founded constructs. By asking generative AI to identify complex constructs of interest, we can potentially alleviate the scarcity of expert labelers. However, relying on proprietary generative AI models introduces challenges, such as dependence on providers, privacy and confidentiality concerns, high costs, and limited reproducibility (Busch 2023; Daniels 2023; Hou and Ji 2024).

To overcome these challenges, we introduce the concept of creating synthetic specialists. Instead of utilizing a generalist AI model trained on vast bodies of knowledge, we develop a specialist capable of identifying a specific complex construct without third-party constraints. We achieve this by approximating powerful AI with an open-source large language model fine-tuned for our classification task of interest. Specifically, we use generative AI to label a training dataset, which is then used to fine-tune the open-source model. Our approach offers greater control, preserves confidentiality, and ensures replicable outputs.

We demonstrate the effectiveness of synthetic specialists empirically by classifying marketing mix elements in consumer-generated content. Our approach addresses the challenges of relying on proprietary generative AI models and the scarcity of expert labelers. We show that synthetic specialists reduce resource

demands and enhance performance in identifying complex constructs, contributing valuable insights for marketers and researchers. By fine-tuning pretrained language models for the classification task, we create specialized models that meet the needs of specific applications without the drawbacks associated with general-purpose generative AI.

Finally, we show that relying solely on popular, but simple metrics like sentiment analysis in marketing research can be imprecise and insufficient for capturing the nuanced perspectives consumers have about brands. Sentiment analysis typically categorizes text as positive, negative, or neutral, but it does not provide insights into the specific aspects of a brand that consumers are discussing (Pang and Lee 2008). By focusing on complex constructs like the marketing mix, we can obtain richer insights into consumer perceptions and experiences.

In our study of over 600,000 tweets about 699 brands, we empirically demonstrate that classifying consumer-generated content using marketing mix elements creates more detailed and actionable insights than sentiment analysis alone. We find that sentiment is distributed differently across marketing mix elements, revealing distinct strengths and weaknesses among competitors. This deeper analysis uncovers specific pain points and areas for improvement in brands' marketing strategies, providing valuable information that sentiment analysis cannot reveal.

GENERATIVE AI: BIGGER, BETTER, OVERKILL?

Generative AI models, such as GPT-4 or Claude 3.5 Sonnet, have gained attention for their wide-ranging capabilities, but they are inefficient when applied to specific tasks like text classification. These general-purpose models require significant computational resources and energy for both training and inference, leading to high costs and a large carbon footprint. Current generative AI models are not only resource-intensive but also pose barriers for marketers lacking access to ultra-high-performance hardware (Desislavov, Martínez-Plumed, and Hernández-Orallo 2023; Kumar and Davenport 2023).

The costs of training these models are staggering and are rarely disclosed by their suppliers. Notable exception are the open-weights llama models by Meta. As shown in Table 1, training Meta's Llama 3.1

405B model required over 16,000 H100 GPUs, consumed an estimated 52 megawatts of electricity, and emitted an estimated 19,346 tons of CO_2. This is equivalent to the annual electricity consumption of 14,979 households. Training took 80 days. Such costs make generative AI impractical for many organizations. See Appendix A for details on the calculations of energy consumption, CO_2 emission, and cost.

Table 1 - Training cost of Generative AI by Example of llama 3.1 405B

Supplier	Meta	**Training H100 GPUs**	16,000	**Training electricity (MW)**	**52,428**	
Model	llama-v3p1-405b-instruct	**Total GPU Hours**	30.840M	**Training CO_2 emissions (tons)**	**19,346**	
Parameters	405 billion	**Training Days**	80	*corresponds to electricity for*	*14,979*	*households*

To mitigate inference costs, models are often pruned or quantized, reducing the precision of their outputs while maintaining sufficient accuracy for most tasks. For example, quantizing Llama 3.1 to INT4 reduces memory and computational requirements, allowing it to run on fewer GPUs. However, even in this optimized state, the high resource demands persist, underscoring the trade-offs between model size, cost, and performance.

While software typically scales well, generative AI is an exception. Running generative AI models for inference (real-time use) requires substantial resources, as seen with Meta's Llama 3.1 model (Table 2). Even when quantized for efficiency, the model still demands high-capacity GPUs and significant energy, producing carbon emissions during every inference task. For example, the H100 GPU, quantized to INT4, consumes 2,500 watts per hour and emits 0.92 kg of CO_2. Such resource demands make using large models untenable for many organizations.

Table 2 - Inference cost of Generative AI by Example of Quantized Llama 3.1 405B (INT4)

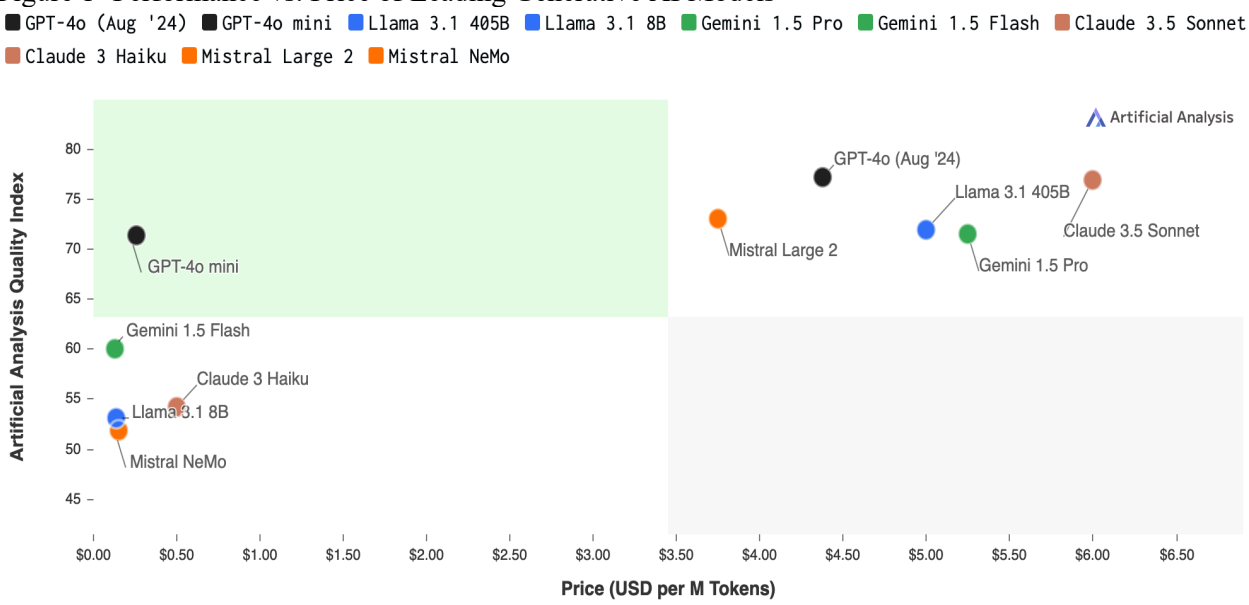
	VRAM (GB)	**GPU** (count)	**Watts** /h	**CO_2** kg/h	**CapEx** GPU invest	**CapEx** /h	**Power** /h	**Own** h	**Rent** /h	**Output Tokens*** /sec
A6000 Ada 48GB	243	6	2,350	0.87	€ 24,273	0.77	€ 0.53	€ 1.30	€ 5.59	14
H100 SXM 80GB	243	4	2,500	0.92	€ 107,880	3.42	€ 0.56	€ 3.99	€ 13.47	28

* *in addition to an initial latency of approximately 1 second*

An important consideration for researchers and managers is whether to rent or purchase the GPUs required for running these models. The capital expenditure (CapEx) for an H100 GPU is approximately

€27,000 (as per September 2024), whereas renting the same hardware through cloud services costs around €3.40 per hour, depending on the provider (we took the mean of Lambda, RunPod, and CoreWeave in September 2024). With a GPU's average lifetime of four years, firms must weigh the initial investment against the ongoing costs of cloud rentals. Over the hardware's lifetime, buying may be more cost-effective for large-scale or continuous operations, but for organizations and individual researchers needing only occasional use, renting offers flexibility without the high upfront investment.

Figure 1- Performance vs. Price of Leading Generative AI Models



Source: www.artificialanalysis.ai; Notes: Artificial Analysis Quality Index: Average result across our evaluations covering different dimensions of model intelligence. Currently includes MMLU, GPQA, Math & HumanEval. OpenAI o1 model figures are preliminary and are based on figures stated by OpenAI.

The performance of generative AI models remains strongly correlated with their size. Larger models tend to generate higher-quality outputs, but their training and inference costs increase exponentially (Kaufmann 2024). As illustrated in Figure 1, the trade-off between performance and price becomes increasingly pronounced as models grow. Notably, Figure 1 suggests that the trend towards larger and more powerful models persists.

In sum, while generative AI models continue to improve in capability, their increasing size and complexity make them costly and inefficient for specific tasks. These inefficiencies suggest that, for many applications, smaller, task-specific models offer a more practical alternative. Synthetic specialists, fine-

tuned for specific tasks, can provide the necessary performance without the massive resource demands associated with general-purpose models.

RELATED LITERATURE

Our research connects the rich literature on text classification with the emerging literature on generative AI in the marketing context. Text classification has long been a vital tool in marketing research for informing strategic decisions. Accurately classifying textual data enables marketers to extract meaningful insights from vast amounts of unstructured information (Berger et al. 2020; Rust et al. 2021). One of the most established methods in this domain is sentiment analysis. This approach categorizes text based on its positive, neutral, or negative tone (Kiritchenko, Zhu, and Mohammad 2014). Recent advancements have expanded sentiment analysis to include paralinguistic cues (Luangrath, Xu, and Wang 2023), measures of certainty (Rocklage et al. 2023), and the fine-tuning of large language models (LLMs) to enhance classifier performance (Hartmann et al. 2023).

Despite its value, sentiment analysis has limitations. It often fails to capture the rich, nuanced information embedded in textual content (Archak, Ghose, and Ipeirotis 2011). To address this gap, researchers have explored methods to attribute sentiment to specific constructs of interest. For example, Chakraborty, Kim, and Sudhir (2022) analyzed restaurant reviews to determine which aspects of the dining experience influenced customer perceptions. By identifying sentiments related to specific service attributes, they provided deeper insights with immediate managerial implications.

Classifying complex constructs in text is a challenging endeavor. Rust et al. (2021) developed and validated 11 language dictionaries to capture dimensions of the value-brand-relationship framework (Rust et al. 2004 Rust, Lemon, and Zeithaml 2004) from Twitter posts. Similarly, Suslava (2021) constructed an elaborate dictionary of corporate euphemisms, augmented with syntactic rules, to study their impact on investor reactions. While dictionary-based approaches offer transparency, they often require significant effort to develop and may not generalize well across contexts.

Supervised machine learning methods have shown promise in improving classification performance for complex constructs (Frankel, Jennings, and Lee 2022; Hartmann et al. 2023). Transformer models like BERT and RoBERTa have been fine-tuned for specific tasks, demonstrating enhanced accuracy. For instance, Puranam, Kadiyali, and Narayan (2021) fine-tuned transformer models on an annotated dataset of 12,000 reviews to analyze service quality perceptions in the restaurant industry. They examined the effect of minimum wage increases by identifying frequency and sentiment related to service attributes.

The scarcity and cost of domain experts pose significant challenges for labeling data in supervised learning. Active learning has been proposed as a solution to optimize the labeling process by selecting the most informative examples (Cohn, Atlas, and Ladner 1994; Chen, Liu, and Wang 2023). However, active learning focuses on which data to label rather than addressing the fundamental issue of labeling complex constructs efficiently.

Generative artificial intelligence offers a potential solution to this problem. Recent studies have explored using generative AI models as substitutes for human experts in various tasks. Horton (2023) demonstrated that LLMs from OpenAI produce responses to economic scenarios consistent with human intuition. Guo et al. (2023) found that ChatGPT-3.5 provides answers similar to those of human experts across multiple domains. These findings suggest that generative AI could surrogate human experts for complex labeling tasks.

In the marketing context, researchers have begun to leverage generative AI for specific applications. Brand, Israeli, and Ngwe (2023) utilized transformer-based models to generate survey responses, finding that the estimated willingness-to-pay derived from AI-generated data was comparable to human studies. They further demonstrated that fine-tuning GPT models with previous survey data improved alignment with human responses for new product features.

Li et al. (2024) investigated the use of LLMs as substitutes for human participants in market research. They generated brand maps capturing customer perceptions and showed that GPT-4 could produce text-based responses closely aligning with human survey results. Castelo et al. (2024) explored the creative capabilities of GPT-4 in generating innovative product ideas. Their study found that GPT-4's ideas were

rated as more creative than those generated by laypeople and professionals, excelling in both form and substance. Reisenbichler et al. (2022) fine-tuned a GPT-2 model to generate search engine optimized content, improving website rankings.

The use of generative AI extends to providing automated feedback in educational contexts. Jürgensmeier and Skiera (2024) employed generative AI to offer students feedback on complex exercises involving coding, statistics, and economic reasoning. Goli and Singh (2024) explored the viability of LLMs like GPT-3.5 and GPT-4 in emulating human survey respondents for eliciting preferences, focusing on intertemporal choices. Arora, Chakraborty, and Nishimura (2024) proposed an AI-human hybrid approach to qualitative and quantitative marketing research. They demonstrated how LLMs could collaborate in the insight generation process, enhancing the efficiency and depth of analysis.

Xia and Liu (2022) provided an initial exploration into classifying brand-authored tweets regarding marketing mix elements. They acknowledged the limitations of their small training sample and called for larger datasets labeled by domain experts. Their work highlights the nascent stage of research on classifying more complex constructs like the marketing mix in consumer-generated content.

Despite these advances, challenges remain in efficiently labeling data for complex classification tasks. Our research builds on this emerging literature by proposing the use of generative AI as a surrogate for expert labelers. By leveraging the vast latent information embedded in generative AI models, we aim to alleviate the scarcity of domain experts and improve classification performance for complex constructs.

Our approach aligns with the call for research on how generative AI can add value to both research and practice (Chui, Roberts, and Yee 2022; Peres et al. 2023; Van Dis et al. 2023). By fine-tuning open-source language models using labels generated by powerful AI, we create specialized classifiers—synthetic specialists—that can efficiently and accurately identify complex constructs without the drawbacks of relying on proprietary models.

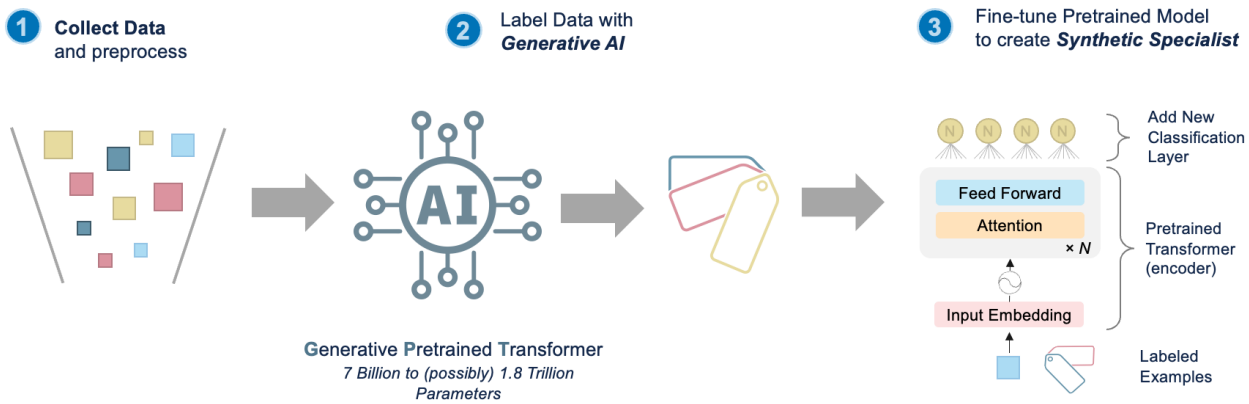
Our work contributes to the literature by demonstrating how synthetic specialists can overcome the limitations of existing methods in text classification. We address the challenges of resource demands, dependency on third-party providers, and the need for extensive expert labeling. By doing so, we offer a

practical and efficient solution for extracting richer insights from consumer-generated content, ultimately enhancing marketing decision-making.

CREATING A SYNTHETIC SPECIALIST

We contend that a specific classification task does not necessitate a massive generative AI model like OpenAI's GPT-4o or Meta's llama 3.1 405B. Instead, marketing analysts typically require an expert model that (1) understands natural language, and (2) can identify specific constructs of interest in text. Fine-tuning a pretrained LLM with examples labeled by generative AI meets both requirements and provides analysts with a highly efficient synthetic specialist that they have full control over. We propose to create synthetic specialists in three phases (see Figure 2).

Figure 2 - Creating a Synthetic Specialist in three Phases



Phase 1: Data Collection

The initial phase focuses on gathering data relevant to the construct of interest. In our study, we collect tweets using the Twitter API. These data represent consumer-generated content mentioning various brands. Preprocessing steps are essential to ensure data quality and relevance. In text analysis, these steps may include removing hyperlinks, personal identifiers, and toxic content. The goal is to assemble a dataset that captures the variability and nuances associated with the construct of interest—here marketing mix elements that consumers' posts pertain to.

Phase 2: Data Labeling with a Synthetic Generalist

After collecting and preprocessing the data, the next step is to label them using a generative AI like GPT-4o. Generative AI models are typically large generative pretrained transformers, commonly comprising 7 billion to 1.8 trillion parameters. They have been trained on vast amounts of data from the internet, including books, articles, and websites. Their training on these data enable them to understand and generate human-like text across a wide range of topics.

The key to utilizing a generative AI model for efficient labeling is zero-shot learning. Zero-shot learning refers to a model's ability to perform a task—such as classifying data into specific categories—without any prior training on that particular task or exposure to labeled examples (Palatucci et al. 2009). Instead, the model relies on the knowledge and patterns it has acquired during its extensive pretraining to make predictions.

In the context of our study, zero-shot learning allows the generative AI to classify Tweets into marketing mix elements (product, price, place, promotion) without having been specifically trained on labeled examples of these categories. This is possible for two reasons: First, the generative AI has been exposed to a vast array of textual data covering numerous topics, including marketing concepts and terminology. It has developed a contextual understanding of how certain words and phrases relate to specific marketing mix elements. Second, the generative AI can comprehend prompts (i.e., instructions) formulated in natural language. By crafting a clear and detailed prompt that defines the classification task, we guide the generative AI to apply its knowledge appropriately.

We use RTF (Role-Task-Format) prompting in this study. An RTF prompt includes instructions on the model's role (e.g., you are a marketing expert), the task (e.g., identify what MMX elements a text pertains to), and the response format (e.g., a list of labels). The generative AI then processes the prompt and the supplied example to generate labels. This process is repeated for all examples in the dataset, enabling automated labeling without manual intervention.

By leveraging zero-shot learning with a synthetic generalist, we efficiently generate high-quality labels for our dataset without the immediate need for human experts. This approach harnesses the power of

advanced AI to overcome traditional barriers in data labeling, setting the stage for creating a specialized model tailored to our specific classification task.

Phase 3: Fine-Tuning to Create a Synthetic Specialist

In the final phase, we transform a smaller, pretrained encoder model—such as BERT (Devlin et al. 2018)—into a task-specific synthetic specialist through fine-tuning. Pretrained encoder models are general-purpose models that provide a universal understanding of language that can be used for a wide variety of tasks (Manning 2022). An abundance of pretrained LLMs, models already trained on various datasets, are available online[1]. These pretrained LLMs can easily be fine-tuned with task-specific training data using transfer learning (Howard and Ruder 2018). Ideally, the pretrained model is trained on text like our target data—in this case, social media content—to ensure it captures the linguistic nuances and context relevant to our dataset.

Using the dataset labeled by the generative AI in Phase 2, we fine-tune the entire model, including both the pretrained layers and the new classification layer. This process is efficient because the pretrained model has already been trained on massive amounts of data, endowing it with a comprehensive understanding of language structures and semantics. We do not need to teach the model about language fundamentals; instead, we adapt it to our specific classification task, which requires substantially less time and computational resources.

During fine-tuning, the model's internal parameters are adjusted to associate input texts with the correct labels by minimizing a loss function, typically cross-entropy loss for multi-class tasks, and binary-cross-entropy for multilabel tasks. The process involves backpropagation, where gradients are calculated, and weights and biases are updated iteratively to improve classification accuracy. This step effectively tailors the general language understanding of the pretrained model to our specific classification task.

Through this fine-tuning process, we create a synthetic specialist that approximates the performance of the generative AI but is optimized for our specific needs. The synthetic specialist is efficient, requiring

[1] See https://huggingface.co/models for a large selection of pretrained LLMs.

significantly fewer computational resources for inference. It can be deployed on standard hardware, such as a personal laptop, and provides reproducible results due to its deterministic nature. By eliminating dependence on third-party models, we gain greater control over the model's behavior and data confidentiality, making the synthetic specialist a practical and effective tool for analyzing complex constructs in textual data.

EMPIRICAL STUDY

The classification task of this study is to determine which of the four marketing mix elements (product, price, place, and promotion) Twitter posts pertain to. We chose the marketing mix because it is at the heart of firms' marketing decisions, and because the construct is sufficiently complex for our investigation (i.e., a multifaceted abstraction). Although central to marketing, we found no publicly available marketing mix classier online. In contrast, there are over 4,000 publicly available sentiment classifiers on www.huggingface.co. We chose Twitter posts because social media continues to be a valuable information source for marketing research (e.g., Liaukonytė, Tuchman, and Zhu 2023; Mallipeddi et al. 2022; Schoenmueller, Netzer, and Stahl 2023), and because their brevity and use of informal language make them a sufficiently challenging baseline for advanced text analysis. By example of the marketing mix, we seek answers to the following four questions:

(R1) Can crowdsourced amateurs correctly identify a complex construct in text?

(R2) Can generative AI correctly identify a complex construct in text?

(R3) Does a task-specific approximation of powerful generative AI perform equally well as the AI?

(R4) Is the disambiguation of consumer sentiment into brands' marketing mix necessary for informed decision-making?

Data Collection

We use tweets collected through Twitter's API in this study. Our data comprise Tweets from 2019 to 2021 that mention major brands' Twitter handle, for example:

@nike in "*Best cushioning ever!!! 🤗🤗🤗 my zoom vomeros are the bomb🏃‍➡️💨!!! @nike #run #training*

We include 699 major brands in our investigation—compiled from the list of brands investigated by Dew, Ansari, and Toubia (2022) and top brands according to YouGov[2].

We built a pool of Tweets by randomly sampling 50 Tweets for each brand and removing duplicate Tweets. We preprocessed all Tweets by removing excessive spaces and breaks, translating HTML entities, and removing URLs. Finally, we drew two mutually exclusive samples from our pool: 1,000 Tweets as validation sample to investigate label agreement among experts, crowdsourced amateurs, and generative AI; 30,000 Tweets as training sample pool from which we train a synthetic specialist.

Humans vs. AI in Text Classification

To assess the viability of surrogating experts with AI in complex classification tasks, we measure how strongly human labels for the four marketing mix elements in the validation sample agree with those of 27 generative AI models by six different suppliers (Anthropic, Google, Meta, Microsoft, Mistral AI, and OpenAI). We measure label agreement using Krippendorff's α to control for chance agreement (Hayes and Krippendorff 2007). Krippendorff's α is the ratio between the observed weighted percent agreement and the chance weighted percent agreement of labels. It ranges from -1 to 1, with 1 representing unanimous agreement, 0 indicating random label assignments, and negative values suggesting systematic disagreement. Labels are considered reliable when $\alpha > .800$, acceptable when $.670 < \alpha < .800$, and insufficient when $\alpha < .670$. Although we chose Krippendorff's α as the main metric for our evaluation, alternative classification metrics including Precision, Recall, and F1-score, and Hamming Loss for multi-label classification (Hamming 1950) exist. Because our findings are consistent across all these metrics (see Appendix B), we only report Krippendorff's α for brevity in the main text of this study.

We started by collecting expert labels for the validation sample. To obtain an objective ground truth and mitigate the risk of label noise (i.e., incorrect labels), four domain experts participated in six workshops to jointly examine, discuss, and label each Tweet of the validation sample. We chose this approach because

[2] https://business.yougov.com/product/brandindex

tweets can be difficult to read and understand, requiring broad knowledge on social media discussions and language. By discussing every tweet individually, experts were able to point different aspects out to the team that other experts may have otherwise overlooked. On average, the team of experts labeled 1.4 Tweets per minute. We use the 4,000 expert labels (1,000 Tweets × 4 labels each) as ground truth in our assessment.

Next, we crowdsourced amateur labels from workers on Amazon's Mechanical Turk (mTurk) platform. Our objective was to investigate whether our labeling task necessitated the use of scarce and costly experts. Workers were briefed on the definition of the four Ps of marketing and tasked to identify which of the four Ps of marketing, if any, a presented Tweet pertains to. To avoid worker fatigue, each worker was limited to labeling 50 Tweets. We used attention checks to ensure higher label quality. We recruited only mTurk Masters (workers with a record of superior performance) and had at least a 90% HIT approval rate (HITs are Human Intelligence Tasks and, in our case, correspond to labeling a single Tweet)[3]. We paid workers an equivalent of USD 12 per hour and obtained a University IRB exemption before fielding the task. Three different workers labeled each Tweet[4]. We used majority votes to obtain a final set of labels for each Tweet.

Finally, we used API from multiple providers[5] for serverless access to 27 different generative AI models. We queried each model on the four Ps of marketing in the validation sample. Specifically, we instructed models with the following role-task-format (RTF) prompt:

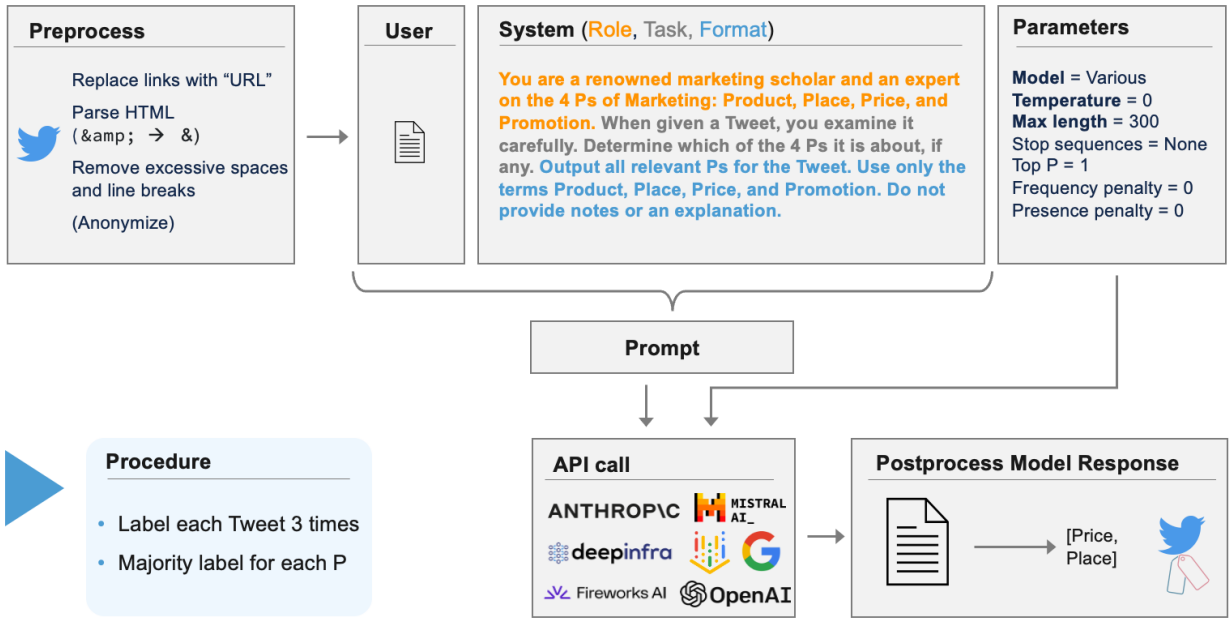
You are a renowned marketing scholar and an expert on the 4 Ps of Marketing: Product, Place, Price, and Promotion. When given a Tweet, you examine it carefully. Determine which of the 4 Ps it is about, if any. Output all relevant Ps for the Tweet. Use only the terms Product, Place, Price, and Promotion. Do not provide notes or an explanation.

Because as generative AI models are autoregressive, we queried the generative AI models for each Tweet individually to avoid self-contextualization where the model considers its own previous output (i.e., label) when labeling the next Tweet. We set models' temperature hyperparameter to zero[6] to obtain

[3] We had a rejection rate of 13.3% due to attention check failures.
[4] In total, 53 mTurk workers participated. Mean labeling time per Tweet was 29 seconds. Total cost for the study was $410.75.
[5] We used the APIs of OpenAI, Anthropic, Google Vertex AI, Mistral AI, Fireworks AI and Deepinfra.
[6] The exception is OpenAI's o1-preview model, that does not allow setting this parameter at the time of writing.

consistent responses (temperature controls response stochasticity; as it approaches zero, the model becomes deterministic). Nonetheless, we found variation in model labels for the same Tweets (see Figure 4). Hence, we followed the same approach as with crowdsourced amateurs: for each generative AI model, we obtained three sets of labels for each Tweet and took the majority vote. See Figure 3 for an overview of the labelling process.

Figure 3 - Labeling Tweets with Generative AI (via API)



Fine-tuning a Pretrained Large Language Model

To create our synthetic specialist on marketing mix elements, we fine-tuned an open-source LLM on generative AI's labels. We proceeded as follows: First, we used generative AI to label the entire training sample of 30,000 tweets three time and take the majority vote. We followed the same procedure as for the validation sample (Figure 3).

Next, we downloaded an open source pretrained encoder model and fine-tune it as a multi-label classifier using our labeled training sample. We selected BERTweet as our pretrained model (Nguyen, Vu, and Nguyen 2020). BERTweet is an encoder model trained on 873 million English tweets using the

RoBERTa procedure for efficiency (Liu et al. 2019). It has 354 million parameters, making it approximately 2,800 times smaller than a generative AI model like GPT-4o[7].

The possibility exists that a Tweet pertains to more than one marketing mix element (i.e., P). Hence, we face a multi-label classification task (a text can pertain to multiple classes), not a multi-class classification task (a text pertains to one of the multiple classes). Consider the following text:

> *@Sony's XM3's ain't as sweet as my bro's airpod pros but got a real steal* 🤑 *the other day* #deal #headphonez.

The writer compares in-ear headphones of two brands (product) and rationalizes purchasing the inferior headphones with a big discount (price). The distinction between multi-label and multi-class classification tasks is conceptually important because it determines the loss function required for fine-tuning the LLM. While multi-class classifiers are typically trained on cross-entropy loss, we used binary-cross-entropy loss in our multi-label classification task because each label corresponds to an independent, binary decision.

For fine-tuning, we used 15,000 tweets randomly drawn from the 30,000-tweet sample pool, with an 80% training and 20% test split. Because later evaluation reveals that no generative AI model achieves reliable agreement (see Figure 5), we pool three generative AI models (GPT-4o, Llama 3.1 405B, Llama 3.1 70B) through majority voting to improve label agreement. Our idea of pooling multiple generative AI models is based on the observation that models performed differently on each of the 4 P's (see Figure 6). By pooling them, we can draw on the relative strengths of each generative AI model.

From a technical perspective, we added a new classification layer to the pretrained model for fine-tuning. This fully connected layer maps the model's internal representations to the four marketing mix categories and adds 4,100 parameters to the pretrained model ($1{,}024 \times 4 + 4$). We used a low learning rate (1 x e-5) and small batch sizes (16) to prevent overfitting. Training was conducted over three epochs with early stopping. The entire fine-tuning process took approximately 20 minutes on a MacBook pro, as we

[7] Based on the assumption that GPT-4o has 1 trillion parameters. Note that OpenAI did not officially disclose model details at the time of writing of this paper.

built upon a model already pretrained on massive amounts of data. This efficiency is achieved because we only need to adapt the pretrained model to our specific task rather than training it language fundamentals from scratch.

Fine-Tuning a Generative AI Model

We also explored fine-tuning a generative AI model. On platforms like OpenAI, it is possible to fine-tune models such as GPT-4o and GPT-4o-mini with domain-specific data and tasks. While inner workings of fine-tuning process are not fully disclosed, it allows for some customization and specialization of generative AI models to specific tasks.

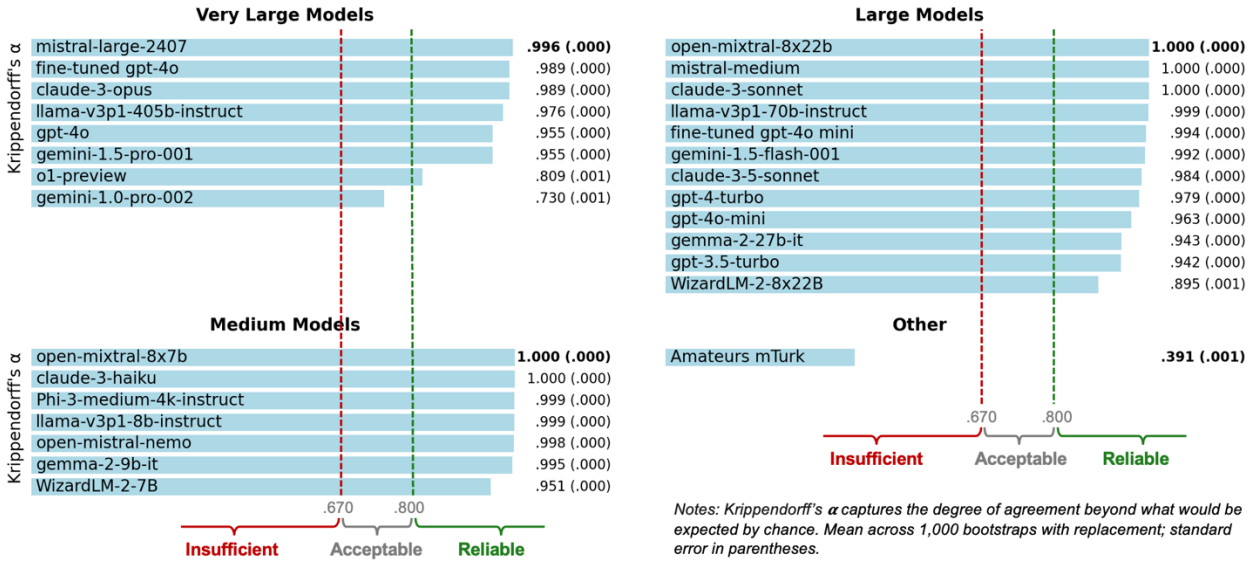
For this study, we manually labeled an additional 200 tweets (not part of the validation or training samples) and used them to fine-tune GPT-4o and GPT-4o mini on OpenAI's platform. The fine-tuning was conducted over four epochs with automatic adjustments of the learning rate. We included these two fine-tuned models in our performance evaluation for comparison purposes.

Nonetheless, we note that fine-tuning proprietary models may improve performance but does not mitigate other shortcomings. These include limited control over capabilities, challenges with reproducibility, higher costs, lower speeds, and the dependency on third-party providers.

Reproducibility of Generative AI Models

Figure 4 presents Krippendorff's α across three runs of 27 generative AI models, grouped by size. To ensure robust findings, we bootstrapped 1,000 times with replacement and report the mean and standard error (in parentheses). We find that very large and large models generally show reliable agreement, with several achieving perfect scores (α=1.000). Medium models also demonstrate high reliability, while crowdsourced amateur labels show insufficient agreement. These results suggest that marketing analysts can choose from several generative AI models to obtain reproducible outcomes. However, the largest and most powerful models should only be used when perfect reproducibility is not required.

Figure 4 - Reproducibility: Label Agreement across three Runs

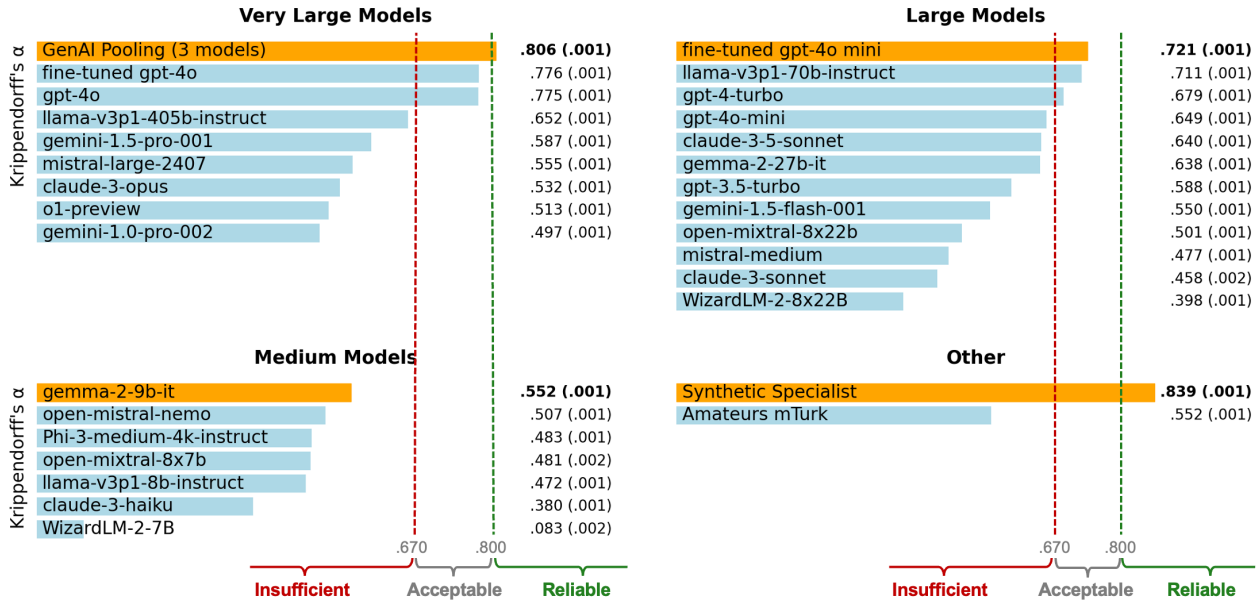


Notes: Krippendorff's α captures the degree of agreement beyond what would be expected by chance. Mean across 1,000 bootstraps with replacement; standard error in parentheses.

Evaluation of Expert Agreement

We present overall expert label agreement (measured with Krippendorff's α) of all models, grouped by model size, in Figure 5. We find that the Synthetic Specialist demonstrates the best overall performance, with a Krippendorff's alpha of .839, placing it in the reliable category. Notably, this performance is 4% better than the GenAI Pooling group (.806), even though the Synthetic Specialist was trained on labels provided by the generative AI models. This difference may be explained by the Synthetic Specialist's fine-tuning on a narrow task, its use of BERTweet (a model specifically trained on Tweets), and its bi-directional encoder architecture, which may be more effective at learning from data and reducing label noise when ample labeled examples are available.

Figure 5 - Agreement with Expert Labels by Model Size



Krippendorff's α: mean across 1,000 bootstraps with replacement; standard error in parentheses. **GenAI Pooling**: GPT-4o, Llama 3.1 405B, Llama 3.1 70B

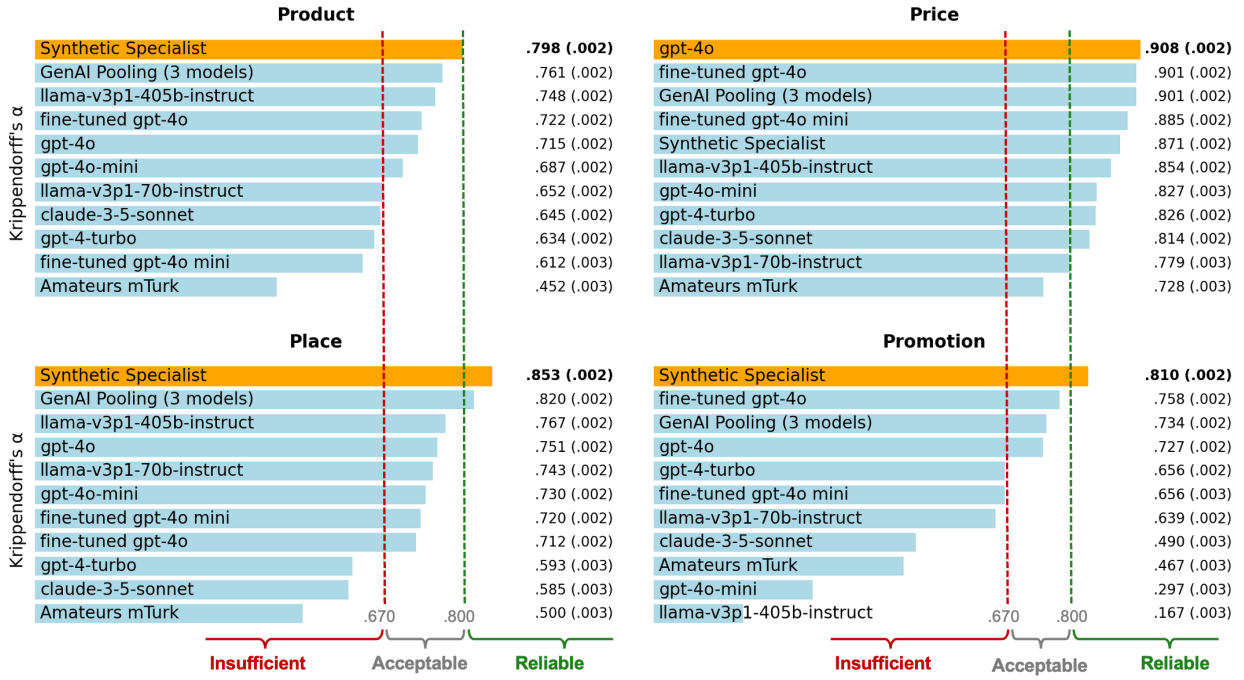
The GenAI Pooling group, which combined three large models (GPT-4o, Llama 3.1 405B, Llama 3.1 70B), also shows reliable agreement but at a potentially high computational cost due to the use of multiple models. Other noteworthy performers include gpt-4o and fine-tuned gpt-4o, along with medium models like Llama 3.1 70B and gpt-4 turbo, which fall in the acceptable category. In contrast, crowdsourced labels from Amateurs (mTurk) have insufficient agreement (.552). The small standard errors across 1,000 bootstraps indicate robustness in these findings, suggesting that the reported agreement levels are stable.

Our findings provide clear answers to our first three research questions. First, crowdsourced amateurs (R1) demonstrate insufficient agreement (Krippendorff's $\alpha = .552$), indicating they struggle to correctly identify complex constructs in text. Second, generative AI models (R2), particularly when pooled, show reliable agreement ($\alpha = .806$), suggesting they can effectively identify complex constructs, though at a high computational cost. Finally, the task-specific approximation of generative AI, the Synthetic Specialist (R3), performs even better than the pooled AI models ($\alpha = .839$), demonstrating that a narrowly fine-tuned model can match or exceed the performance of more powerful, general-purpose generative AI models when tailored to a specific task.

Expert Agreement by Marketing Mix Element

Examining each marketing mix element individually is important because models may perform differently across elements, and overall performance can mask these variations. Figure 6 presents the top 10 models for each marketing mix element. We find that model performance is not uniform across product, price, place, and promotion.

Figure 6 - Expert Agreement by Marketing Mix Element: Top 10 Models



Krippendorff's α: mean across 1,000 bootstraps with replacement; standard error in parentheses. **GenAI Pooling**: GPT-4o, Llama 3.1 405B, Llama 3.1 70B

The Synthetic Specialist shows reliable agreement on three elements but falls just short on product ($\alpha = .798$). Llama 3.1 405B performs reliably on price and acceptably on product and place but struggles with promotion, where its performance is insufficient. Pooling models, as proposed in this study, can achieve a more balanced and reliable outcome across marketing mix elements. We combined GPT-4o for its overall acceptable performance with Llama 3.1 405B (which excels in product and place but struggles with promotion) and Llama 3.1 70B (which consistently ranks in the top 10 across all elements).

We also find that fine-tuning GPT-4o leads to only slight improvements, whereas fine-tuning GPT-4o-mini produces more substantial gains. This may be because the smaller model (GPT-4o-mini) requires fewer fine-tuning examples to adapt effectively. However, without a deeper understanding of OpenAI's proprietary technology, any conclusion would be speculative.

Training Sample Size and Robustness

Model performance can vary depending on the training samples used. We fine-tuned BERTweet five times with different seeds for randomly drawn training samples of varying sizes. Our objectives were twofold: first, to ensure robust performance of our Synthetic Specialist; second, to determine how many labeled examples are necessary to achieve reliable labels.

We evaluated each fine-tuned model on our validation sample using three metrics: Krippendorff's α, area under the receiver operating characteristic curve (AUC), and Hamming loss. Krippendorff's α controls for chance agreement but is fixed to a particular probability threshold (here, 0.5). In contrast, AUC evaluates a model's ability to distinguish between classes regardless of the probability threshold. Hamming loss, commonly used for multi-label classification tasks, considers all labels simultaneously, providing a holistic view of the model's overall error rate. We transformed Hamming loss (1 - Hamming loss) to align it with the other metrics, where higher values are better.

Figure 7 - Label Agreement at different Training Samples and Sample Sizes

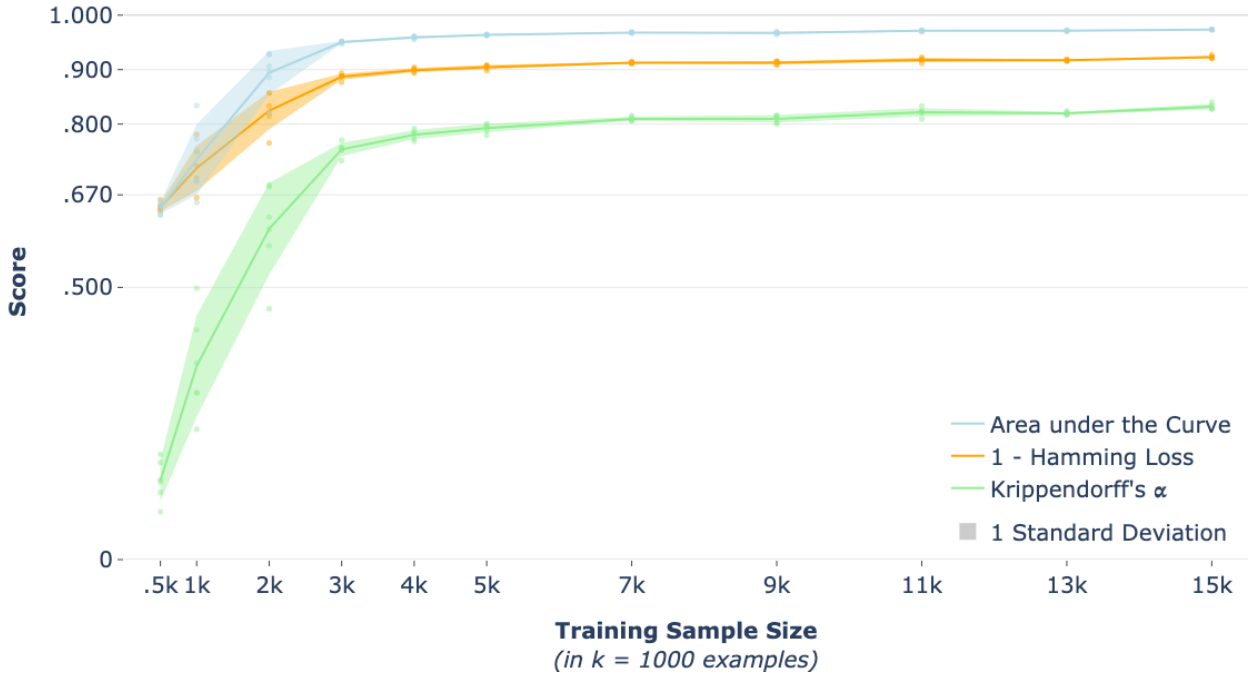


Figure 7 shows that our Synthetic Specialists perform well in identifying which of the four marketing mix elements a tweet pertains to. We observe a very high AUC of .950 with as few as 3,000 training examples. This finding aligns with Hartmann et al. (2023), who found that relatively few training examples can produce good results when fine-tuning large language models. We observe a similar pattern for Hamming loss, which drops substantially up to 4,000 training examples, after which improvements are smaller.

Krippendorff's α rises sharply as the number of training samples approaches 4,000. We achieve reliable agreement from a training set size of 7,000 examples onward. The variance for small training sample sizes is substantial (shaded areas in Figure 7) but becomes minimal from 3,000 training examples onward, suggesting that training sample composition and model initialization are less relevant for larger training samples.

Model Efficiency

Our Synthetic Specialist outperforms the next best approach—GenAI Pooling—by 4%. We compare the efficiency of the two approaches in Table 6. The Synthetic Specialist labels tweets 297 times faster than the

pooled GenAI approach. The cost is dramatically lower, at just 0.007% of the cost, and with a fraction of the CO_2 emissions. We conclude that Synthetic Specialists may not only be more accurate but are vastly more efficient in terms of speed, cost, and environmental impact.

Table 6. Comparison between Model Efficiency for overall two best models

GenAI Pooling: 3 models x 3 runs x 1000 Tweets					
	Tweets /sec	Tweets /h	Time* (minutes)	Cost API	CO_2 Emission** (kg)
Fireworks / OpenAI API	.38	1383	43.40	€ 3.56	2.000

** concurrent API calls for models: llama 3.1 405B and 70B, GPT-4o*
*** assuming additional .2 watt per internet query and 50-watt laptop*

Synthetic Specialist: 1 x 1000 Tweets					
	Tweets /sec	Tweets /h	Time (minutes)	Cost on MacBook	CO_2 Emission (kg)
Local Laptop Computer	114.03	410,490	.15	€ .00	.0001
Cost for initial 15k genAI labels (pooled models)			*650.93*	*€ 48.03*	*30.04*

Note: It takes approximately 22min to fine-tune BERTweet with 15k Tweets at 100-watt MacBook Pro

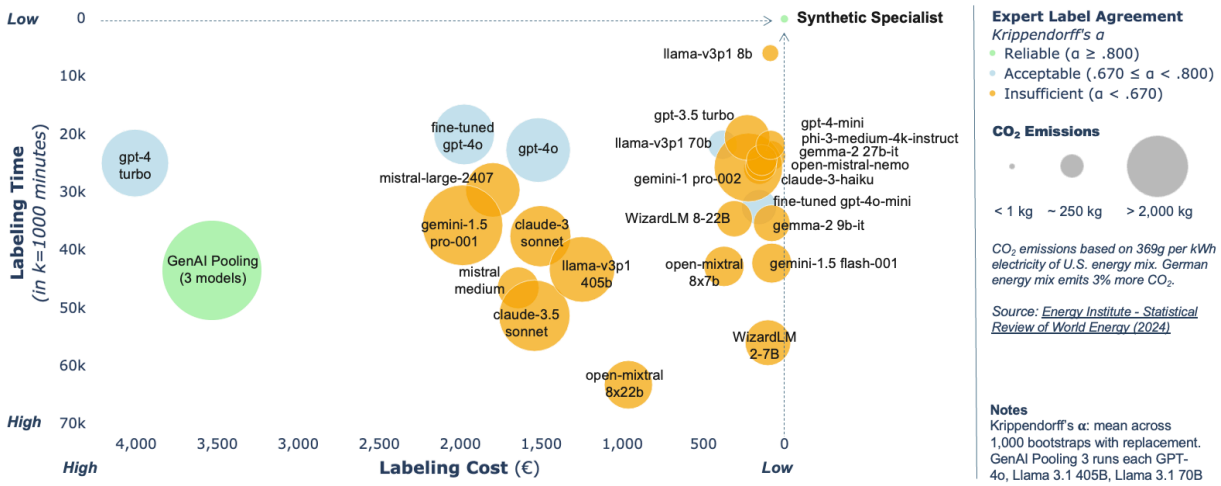
It is important to note that there is an initial one-time cost to label data using pooled generative AI models, which is necessary to train the Synthetic Specialist. This includes both financial cost and emissions from generating the required 15,000 labels. However, once trained, the Synthetic Specialist scales efficiently with negligible additional costs and emissions. In the next section, we demonstrate this scalability through a model comparison across all 27 generative AI models, the pooled model, and the Synthetic Specialist for the classification of 1 million tweets.

Classification at Scale

Classifying large volumes of data entails different costs and considerations for various stakeholders. Analysts are primarily concerned with the time required to label data, as it affects productivity and the speed at which insights can be generated. Firms focus on the financial cost of labeling, measured in euros, since it impacts operational budgets and resource allocation. Society, on the other hand, is affected by the environmental cost, particularly CO_2 emissions resulting from the computational resources used in labeling tasks. To assess these costs, we compare the performance of various generative AI models and the Synthetic Specialist in labeling one million tweets. Figure 8 illustrates this comparison, presenting labeling time (x-

axis), labeling cost (y-axis), and CO_2 emissions (bubble size), along with expert label agreement measured by Krippendorff's α (bubble color).

Figure 8 - Model Efficiency, Effectiveness, and CO_2 Emissions for labeling 1 million Tweets



Labeling time varies significantly across models. Smaller models, such as Llama-v3p1 8B, process data quickly but yield low agreement with expert labels. Larger models, including GPT-4o and GPT-4 Turbo, require substantially more time due their complexity and computational demands. The GenAI Pooling approach, which combines multiple large models, takes over 30 days to label the dataset. In contrast, the Synthetic Specialist labels one million tweets in less than 2.5 hours, demonstrating superior efficiency from the analyst's perspective.

Financial costs also vary heavily among models. Smaller models incur lower costs but provide insufficient label accuracy. Larger models offer better accuracy but at significantly higher expenses due to API usage fees and computational resources. For instance, labeling 1 million tweets with GPT-4o would cost approximately €1,500. Although reliable in terms of expert agreement, the GenAI Pooling approach is among most expensive options. The Synthetic Specialist, however, labels the data at a negligible cost after the initial training investment, making it the most cost-effective solution for firms handling large-scale labeling tasks. Note that we excluded OpenAI's newest o1-preview model (at the time of writing) because at cost of over €120,000, it would have not fit in Figure 8.

The environmental impact, measured in CO_2 emissions, is a critical consideration for society. Larger models consume more energy, leading to higher emissions. The GenAI Pooling approach results in CO_2 emissions exceeding 2 metric tons. In contrast, the Synthetic Specialist emits approximately 100 grams of CO_2 when labeling one million tweets. This stark difference clearly demonstrates the environmental benefits of using efficient, task-specific models over large, general-purpose generative AI models.

Expert label agreement, assessed using Krippendorff's α and shown by bubble color, reveals a trade-off between model size, efficiency, cost, and label accuracy. Most models fall into the insufficient agreement category (orange bubbles), particularly smaller models with lower computational requirements. Medium-sized models, such as Llama-v3p1 70B and fine-tuned GPT-4o Mini, achieve acceptable agreement levels (blue bubbles). Only two approaches—the GenAI Pooling and the Synthetic Specialist—reach reliable agreement (green bubbles) with Krippendorff's α exceeding .800. However, the Synthetic Specialist attains this level of accuracy with significantly lower time, cost, and environmental impact.

Our comparison shows that smaller model, while efficient and inexpensive, fail to provide accurate labels. Larger models deliver better accuracy but are inefficient and costly. The Synthetic Specialist outperforms other models by offering reliable label accuracy while being highly efficient and cost-effective. It minimizes the time analysts spend on labeling, reduces financial costs for firms, and aligns with societal goals for environmental sustainability. Scaling up the classification task to one million tweets shows the practical advantages of the Synthetic Specialist. It provides reliable label accuracy comparable to that of the most powerful generative AI models but with substantially lower costs in terms of time, money, and environmental impact. For stakeholders seeking to process large volumes of textual data efficiently, the Synthetic Specialist presents a viable solution that benefits analysts, firms, and society alike.

CREATING RICHER MARKETING INSIGHTS FROM USER GENERATED CONTENT

Thus far, we have addressed three research questions. First, human amateurs struggle to identify complex constructs in text accurately. Second, generative AI models can effectively perform this task. Third, a task-

specific approximation of generative AI—the Synthetic Specialist—can match or even surpass the performance of general-purpose AI models while being more efficient. We now turn to our fourth research question: (R4) Is the disambiguation of consumer sentiment into brands' marketing mix elements necessary? To answer this question, we examine whether analyzing sentiment at the level of individual marketing mix elements provides richer insights than overall sentiment analysis.

Our objectives are fourfold. First, we investigate the relationship between overall brand sentiment and sentiment associated with each marketing mix element. Second, we analyze differences in marketing mix sentiment among competing brands and within brand portfolios. Third, we discover underlying topics for a pain point of a major apparel brand that we identified in its marketing mix. Fourth, we extract actionable insights from user-generated content on Twitter based on these topics.

By applying the Synthetic Specialist to a large dataset of tweets mentioning 699 brands, we classify each tweet according to the marketing mix elements: product, price, place, and promotion. We then analyze the sentiment distribution within each category. Our approach allows us to uncover nuanced consumer perceptions that are not apparent from overall sentiment analysis alone.

Consumer Sentiment by Marketing Mix Element

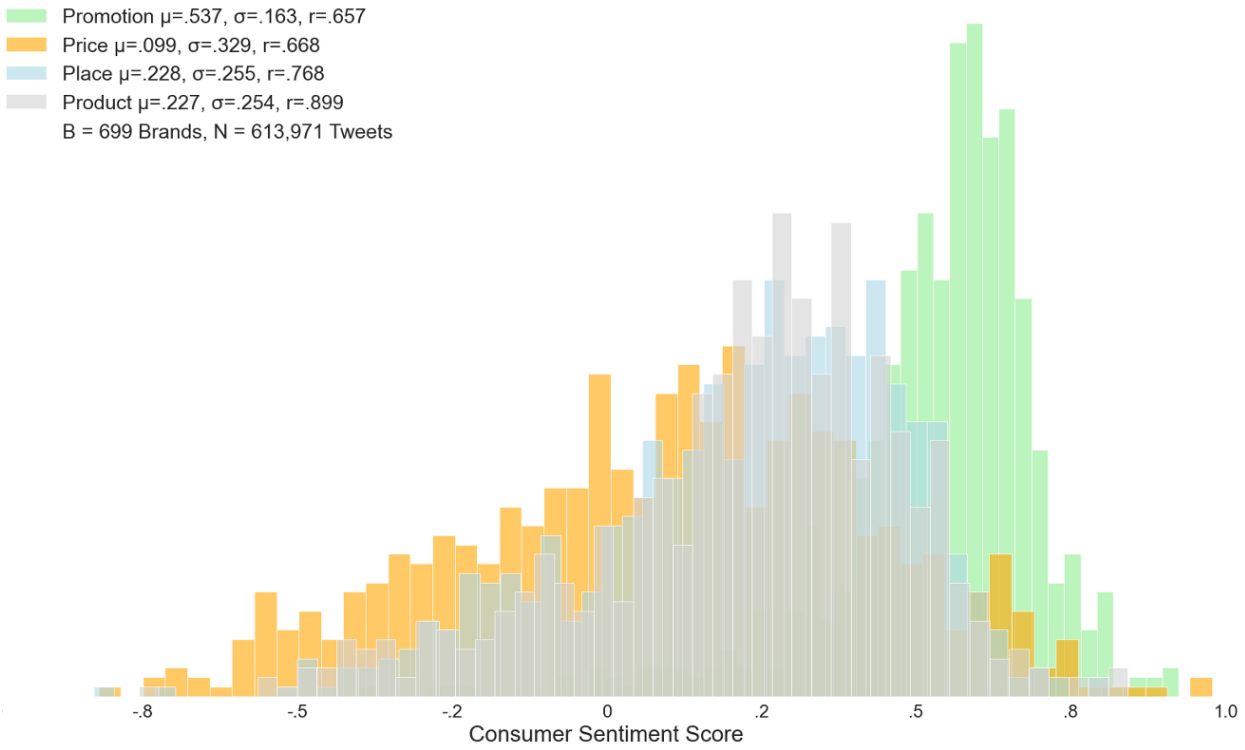
To investigate the relationship between overall brand sentiment and sentiment associated with each marketing mix element, we classified tweets mentioning 699 brands from the year 2020 with our synthetic specialist. For each brand, we randomly sampled up to 1,000 tweets posted in 2020, resulting in a dataset of 613,971 tweets.

We used a fine-tuned RoBERTa model for sentiment analysis, specifically "Twitter-RoBERTa-base for Sentiment Analysis" that we downloaded from the Hugging Face model hub (https://huggingface.co/). It was fine-tuned on approximately 124 million tweets posted between January 2018 and December 2021 (Camacho-Collados et al. 2022). It outputs probabilities for positive and negative sentiment. To obtain a continuous sentiment score ranging from -1 (completely negative) to 1 (completely positive), we calculated a weighted sentiment score as follows:

$$Sentiment\ Score = P(positive) - P(negative) \quad \text{(eq 1)}$$

Figure 9 shows the distribution of consumer sentiment scores by marketing mix elements. Each category exhibits distinct characteristics in its sentiment distribution. We find that promotion has the most positive sentiment, with its distribution centered around a relatively high mean (μ = .537) and low dispersion (σ = .163). Across brands, consumers responses to promotional activities seem are positive. In contrast, price exhibits a much lower mean sentiment (μ = .099) as well as the largest standard deviation (σ = .329). This suggests that consumer sentiment about price is more varied, with a significant proportion of tweets being neutral or negative. The wider spread reflects greater consumer sensitivity and mixed opinions regarding pricing and perceived value.

Figure 9 -Sentiment Distributions by Marketing Mix Element



We observe a more moderate mean sentiment for place and product (μ= .228 and μ = .227, respectively) with less variation than price but more than promotion. (σ= .255 and .254σ, respectively). Apparently, consumers' sentiments about distribution channels or location aspects are more consistent but generally less favorable than for promotions. The high correlation coefficient for product (r = .899) suggests that product-related tweets strongly influence overall sentiment. This dominance could mask consumer

sentiment on brands' other marketing mix elements, possibly misleading marketers about their effectiveness.

Variability in Sentiment across Marketing Elements

Building on our analysis of sentiment distributions by marketing mix elements, we now examine the variability of consumer sentiment within individual brands across these elements. Our objective is to determine the extent to which sentiment differs among the four marketing mix elements—product, price, place, and promotion—for each brand. We calculated the sentiment range for each brand by finding the difference between the highest and lowest sentiment scores among the marketing mix elements (range = maximum sentiment - minimum sentiment). The approach allows us to assess the distribution of sentiment ranges across all 699 brands in our dataset in Figure 10.

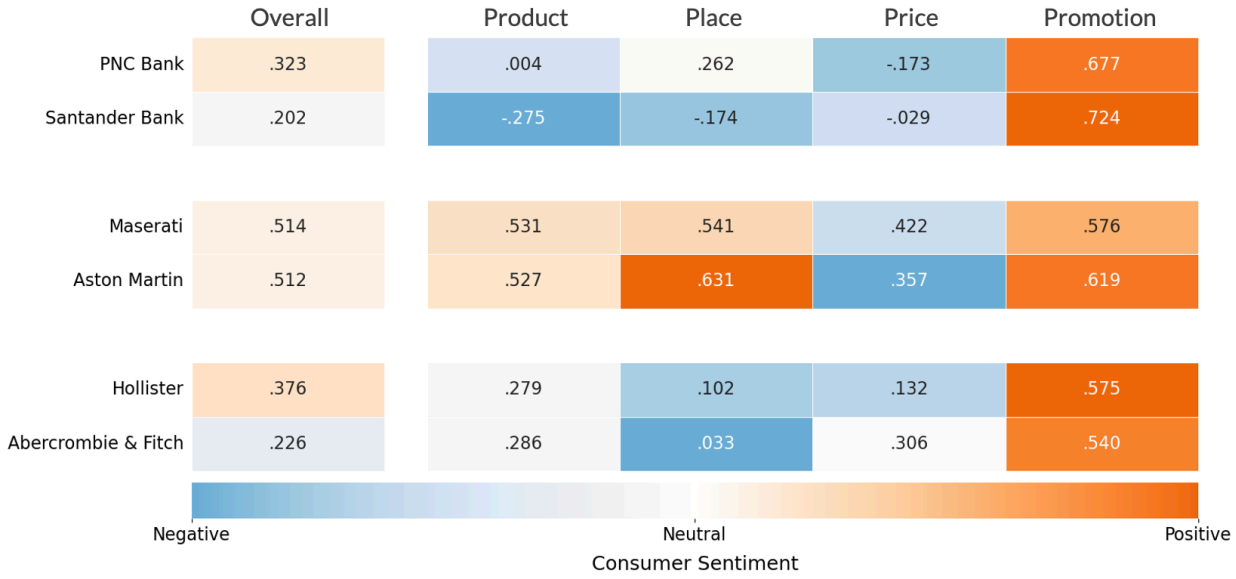
Figure 10 - Variability in Brand Sentiment as Range from lowest to highest Sentiment Score

The box plot in Figure 10 shows an interquartile range from approximately .3 to .7, indicating that half of the brands have moderate variability in their sentiment scores across their marketing mix. The whiskers extend up to about 0.9, encompassing most of the distribution, while several outliers exceed a sentiment range of 1.0, with some reaching as high as 1.8. This substantial variability supports our notion of the necessity for marketers to evaluate each marketing mix element individually (research question 4). Relying solely on overall sentiment scores can mask critical weaknesses or strengths in specific marketing mix elements, potentially leading to misconceptions about marketing mix effectiveness and misguided strategic decisions. By identifying these disparities, marketers can discover brands' unique strengths and weaknesses within the marketing mix and make targeted adjustments to improve customer experiences where it is most needed.

Marketing Mix Insights for Brand Managers

The differences in sentiment distributions and ranges across brands suggest that marketers can gain additional insights when they focus the analysis of consumer sentiment on individual marketing mix elements. We examine three brand pairs in Figure 11 to illustrate how a differentiated view on consumer sentiment can give brand managers deeper insights into their marketing mix and that of their competitors.

Figure 11 - Brand-specific MMX Elements requiring Managers'Attention



Our first example shows that competing brands within the same industry can have different strengths and weaknesses that are not apparent from overall sentiment scores. PNC Bank and Santander Bank are both major financial institutions, but their consumer sentiment profiles differ significantly. PNC Bank shows neutral to positive sentiment across all Ps except price (.173), indicating that while PNC's overall efforts are effective, customers are dissatisfied with its pricing. Santander Bank, on the other hand, has negative sentiment in product (-.275) and place (-.174) but performs best in promotion (.724), suggesting that Santander needs to improve its product offerings and distribution channels. These differences highlight that even firms offering near identical products and services will do so with varying success across the marketing mix, leading to very different positions in consumers' minds.

Our second example shows that brands with similar overall sentiment scores may require different strategic adjustments due to variations in consumer perceptions of marketing mix elements. Maserati and

Aston Martin, both luxury automobile brands with nearly identical overall sentiment (.514 and .512, respectively), differ in how consumers view specific elements. Aston Martin has higher sentiment in place (.631) and product (.527), suggesting that consumers appreciate its distribution channels and product features. Maserati has slightly stronger sentiment in price (.422) but lower sentiment in place and product. This implies that while Maserati may offer better perceived value, it needs to improve its product features and accessibility to meet consumer expectations. Despite similar overall sentiment, these brands must prioritize different areas to enhance their positioning in the luxury car market.

Our third example shows that examining brands within the same corporate portfolio can reveal opportunities for strategic alignment. Abercrombie & Fitch (A&F) and its subsidiary Hollister show notable differences in consumer sentiment. Hollister has a higher overall sentiment score (.376) compared to A&F (.226) and outperforms in promotion (.575 vs. .540) and place (.102 vs. .033). The finding suggests that Hollister is more effectively positioned in terms of promotional strategies and distribution channels. A&F may benefit from adopting some of Hollister's successful tactics on these marketing mix elements. On the flip side, Hollister falls short on price relative to A&F (.132 vs. .306), suggesting that consumers perceive to get more value for their money when purchasing A&F products. Identifying such internal disparities helps brand managers pinpoint opportunities for leveraging portfolio strengths to remedy.

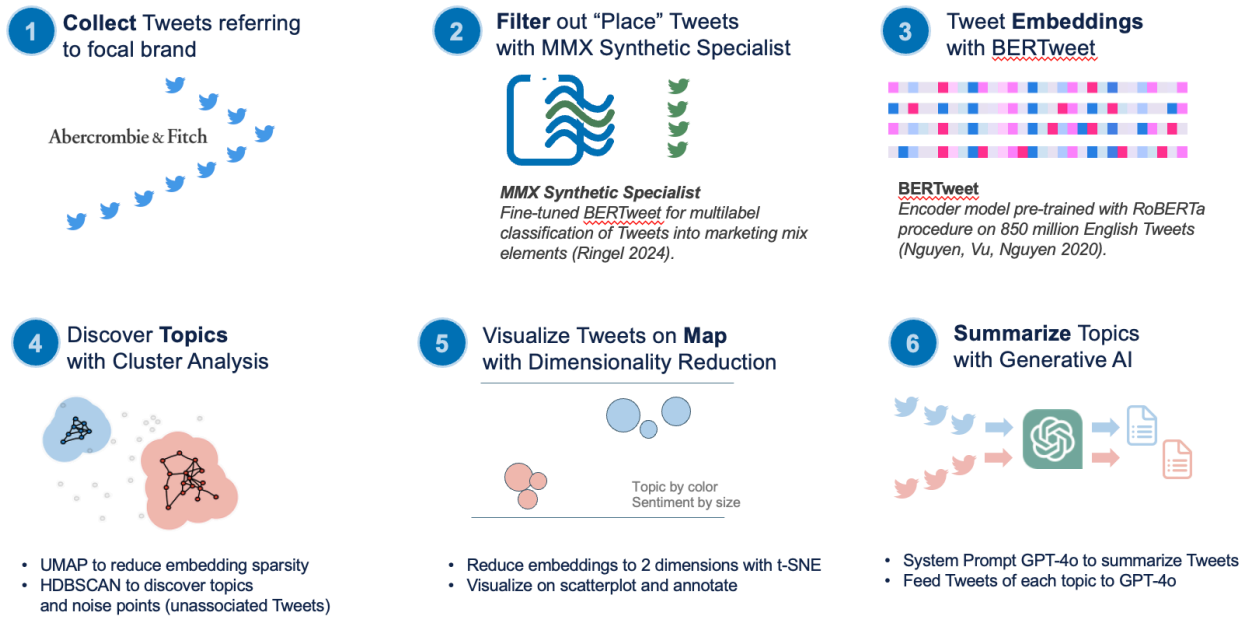
These examples illustrate the importance of analyzing consumer sentiment at the level of individual marketing mix elements. Relying solely on overall sentiment scores can mask critical strengths and weaknesses. Brands with similar overall sentiment may face different challenges and opportunities. By examining sentiment in detail, brand managers can identify specific areas for improvement and tailor their strategies to address consumer perceptions more effectively.

Zooming-in on Abercrombie's biggest Pain Point: "Place"

While sentiment scores at the marketing mix element level provide valuable insights, they do not reveal why consumers' sentiment is high or low. To adjust their marketing mix, brand managers need to understand the root causes of both positive and negative customers perceptions. To illustrate how marketers can obtain these deeper insights, we focus on Abercrombie & Fitch (A&F) tweets related to "Place", which exhibits

the lowest consumer sentiment across A&F's marketing mix. Our objective is to rapidly identify the drivers of consumers' negative sentiment within "Place" and determine the most critical issues that need to be addressed, as well as any positive aspects that can be leveraged.

Figure 12 - Overview of the Approach for Zooming-in on Pain Points



We proceeded using a six-step approach (Figure 12). First, we collected tweets from 2020 mentioning A&F. We then classified these tweets with our synthetic specialist to filter out those pertaining to "Place". This filtering allowed us to zoom-in on tweets that talk about A&F's distribution channels and locations. Next, we generated tweet embeddings using BERTweet, an encoder model pre-trained on 850 million English tweets (Nguyen, Vu, and Nguyen 2020). Transforming the textual data into numerical embeddings captures the semantic relationships between words and phrases, which is essential for clustering similar tweets based on their content.

We conducted cluster analysis in step 4 to discover topics of semantically similar tweets. To facilitate the clustering process, we reduced the dimensionality of the embeddings from 1,024 to 30 dimensions using Uniform Manifold Approximation and Projection (UMAP) by McInnes, Healy, and Melville (2018). High-dimensional data like our tweet embeddings can suffer from the "curse of dimensionality," where the distance metrics become less meaningful due to sparsity and noise in high-dimensional space (Aggarwal,

Hinneburg, and Keim 2001). By applying dimensionality reduction with UMAP, we preserved the semantic relationships while mitigating noise and sparsity, making the subsequent clustering more effective (McInnes, Healy, and Melville 2018). This pre-processing of the embeddings allowed us to better capture the intrinsic structure of the data, leading to more meaningful and interpretable clusters.

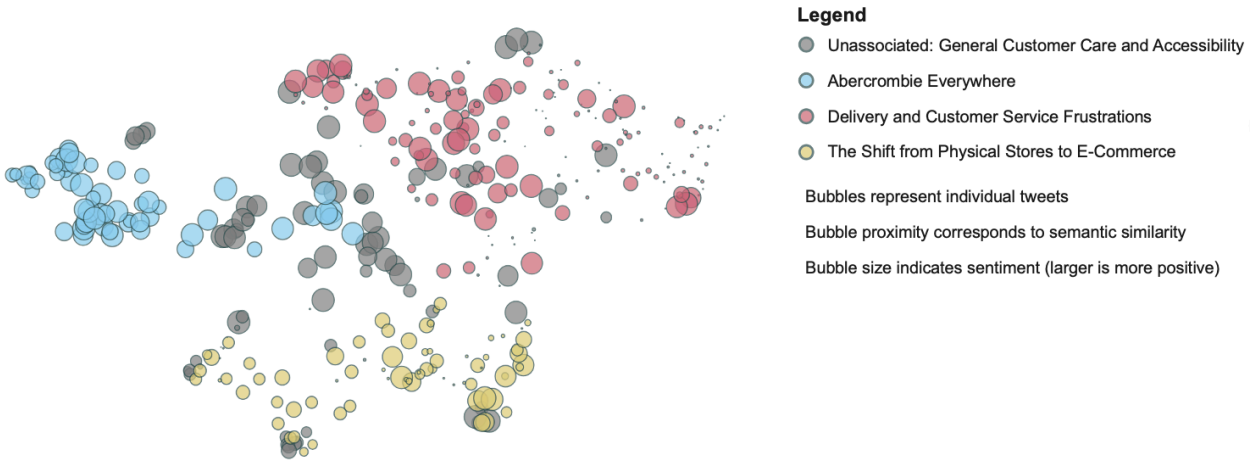
We then used Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to identify clusters and noise points (Campello, Moulavi, and Sander 2013). We chose HDBSCAN because it is non-exhaustive; that is, it does not require every tweet to be part of a cluster. Tweets that cannot be easily associated with any cluster are labeled as noise points. We treated each cluster as a topic of semantically related tweets and manually labeled these topics by examining several tweets in each topic.

We then visualized the clustered tweets on a map using t-distributed stochastic neighbor embedding (t-SNE) by van der Maaten and Hinton (2008). The map (Figure 13) shows tweets as bubbles and is augmented by topic membership (bubble color) and sentiment (bubble size), providing brand managers with an intuitive overview of topic structure and sentiment distribution across topics that they can then zoom-in on further. Finally, we summarized the tweets within each topic using a generative AI model (OpenAI's GPT-4o). Specifically, we provided the generative AI with the tweets of each topic and prompted it to generate succinct summaries focusing on the main themes mentioned by consumers. Utilizing generative AI in this step leverages its capability to synthesize and condense large volumes of textual data efficiently. Such summarization at scale saves brand managers the time-consuming task of reading and manually summarizing hundreds of tweets.

We present the consumer sentiment map of A&F's "Place" tweets in Figure 13. The map shows that tweets cluster naturally into distinct topics, with bubbles of the same color grouped together. Notably, the "Delivery and Customer Service Frustrations" cluster (red) dominates a significant portion of the map and is characterized by smaller bubbles, indicating strong negative sentiment. These small, red bubbles suggest that issues with delivery and customer service are major pain points for consumers. In contrast, the

“Abercrombie Everywhere” cluster (light blue) consists of larger bubbles, reflecting positive sentiment. Consumers express satisfaction with the brand’s global availability and shopping experiences.

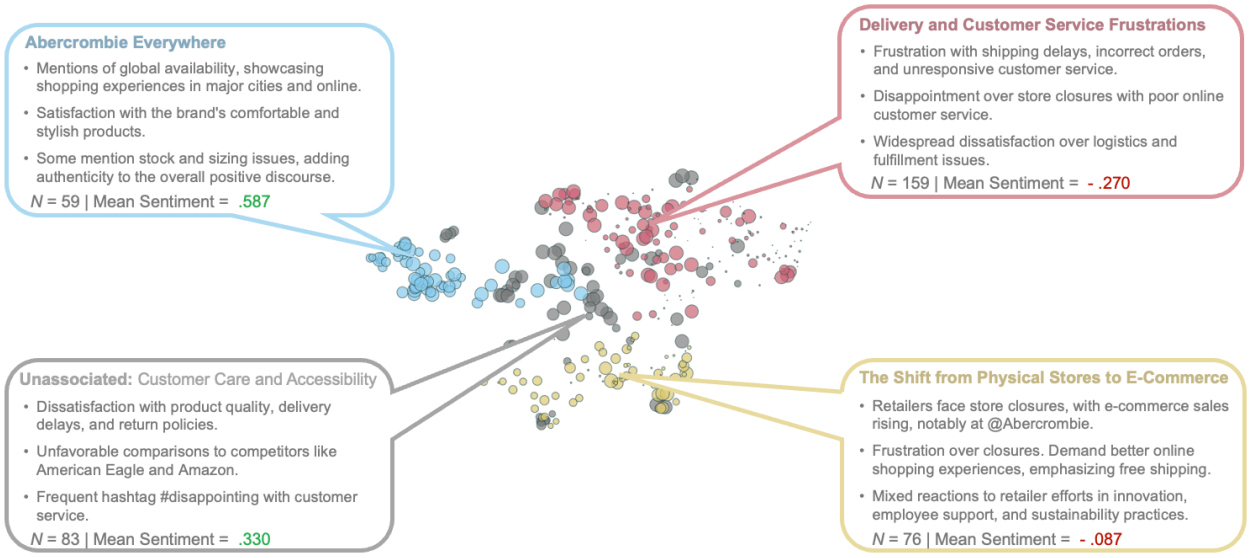
Figure 13 - Map of Abercrombie & Fitch “Place” Pain Points



By augmenting the map in Figure 13 with topic summaries by a generative AI in Figure 14, brand managers can quickly develop a deeper understanding of the specific issues affecting consumer sentiment. We find that the “Delivery and Customer Service Frustrations” topic is the largest, comprising 159 tweets with a mean sentiment of -.270. The topic highlights significant pain points related to shipping delays, incorrect orders, and unresponsive customer service. The prevalence of negative experiences in this area suggests that A&F should prioritize improvements in logistics and customer support to improve customer’s experience.

Conversely, the “Abercrombie Everywhere” topic comprises 59 tweets with a mean sentiment of .587, reflecting positive consumer experiences. Customers express satisfaction with the brand’s global availability, comfortable and stylish products, and enjoyable shopping experiences both in major cities and online. The positive feedback indicates that A&F’s efforts in expanding its presence and offering appealing products are resonating with consumers. Brand Managers can leverage these strengths by emphasizing successful strategies in promotion and accessibility.

Figure 14 - Inside Pain Point Topics with Generative AI



The "Shift from Physical Stores to E-Commerce" topic, with 76 tweets and a mean sentiment of -.087, reveals slight negativity surrounding the transition to online shopping. Consumers express frustration over store closures and demand better online shopping experiences, including free shipping. This indicates a need for A&F to enhance its e-commerce platform and address consumer concerns about accessibility and convenience in the digital marketplace.

Despite "Place" being A&F's greatest pain point, our in-depth analysis reveals that consumer sentiment is not equally distributed within the marketing mix element. By applying our six-step approach, we also uncovered positive aspects such as the "Abercrombie Everywhere" cluster, where consumers express satisfaction with the brand's global availability and shopping experiences. This level of granularity provides brand managers with nuanced insights, enabling them to make informed marketing mix decisions that address specific weaknesses while leveraging existing strengths.

MANAGERIAL IMPLICATIONS

This study presents several implications for marketers and organizations. First, relying solely on overall sentiment analysis is insufficient for informed marketing mix decisions. Our findings show that sentiment is distributed differently across the marketing mix elements—product, price, place, and promotion (Figure

9). Analyzing consumer sentiment at the level of individual marketing mix elements allows brand managers to identify specific strengths and weaknesses. For instance, while Abercrombie & Fitch exhibited negative sentiment in "Place," further analysis revealed both negative and positive aspects within this element (Figures 13 and 14).

Second, the MMX synthetic specialist of this study offers a practical tool for disentangling sentiment into relevant marketing mix elements. Compared to generative AI models like GPT-4o, the synthetic specialist is more efficient and accessible. It labels tweets 297 times faster, costs only 0.007% as much, and emits just 0.004% of the CO_2 emissions (Table 6). These efficiencies make it feasible for organizations to process large volumes of consumer-generated content without significant resource investments.

Third, for organizations, the use of synthetic specialists mitigates dependency on third-party providers and addresses privacy and confidentiality concerns. Since synthetic specialists run locally, firms maintain full control over their data, ensuring compliance with internal policies and regulations. Additionally, synthetic specialists can be updated (through further fine-tuning) or retrained as needed, providing flexibility in adapting to changing market conditions or new classification tasks. Importantly, synthetic specialists' outputs are replicable, which ensures reproducibility of analyses based on their classifications.

Fourth, the minimal hardware requirements of synthetic specialists make advanced analytical capabilities more accessible, potentially benefiting smaller organizations or those with limited resources. By reducing barriers to entry, synthetic specialists promote broader adoption of sophisticated text analysis methods. Moreover, the reduced environmental footprint aligns with organizational sustainability goals.

Fifth, the six-step approach we outlined provides a method for managers to gain deeper insights into consumer perceptions. By clustering and summarizing tweets related to specific marketing mix elements, managers can quickly identify the root causes of consumer sentiment (Figures 12 and 13). The proposed approach facilitates targeted interventions to address negative perceptions and reinforce positive ones. For example, identifying that customer frustrations are concentrated around delivery and customer service at Abercrombie & Fitch helps its managers to prioritize improvements in these areas.

Finally, synthetic specialists can be extended to other constructs and domains. Marketers may apply similar approaches to analyze service quality dimensions, customer experience factors, or branding elements. Our approach may also benefit other fields where complex constructs need to be identified in textual data, such as legal document analysis or policy evaluation.

CONCLUSION

We introduce the concept of synthetic specialists—task-specific approximations of powerful generative AI models. Our approach offers a practical solution for complex classification tasks, bridging the gap between the efficiency of specialized models and the expansive knowledge embedded in generative AI. By focusing on a specific task, synthetic specialists mitigate the drawbacks associated with proprietary AI models, such as high computational costs and dependency on third-party providers.

Empirical results demonstrate that a synthetic specialist can outperforms 27 generative AI models in classifying a complex construct; here, marketing mix elements that tweets pertain to. With 4% higher expert agreement, our marketing mix synthetic specialist not only achieves superior accuracy but also operates with significantly greater efficiency. It labels data 297 times faster than pooled generative AI models, costs only 0.007% as much, and emits a fraction of the CO_2 emissions. These findings highlight the synthetic specialist's effectiveness and sustainability for large-scale text classification tasks. Notably, our empirical evaluation shows that the use of generative AI for reliable MMX classification at scale is not viable: labeling 1 million tweets would take approximately 30 days, cost over € 3,500, and produce two tons of CO_2.

The implications for marketing research and practice are substantial. By disaggregating consumer sentiment into individual marketing mix elements, researchers and practitioners obtain nuanced insights into consumer perceptions. Such more granular analysis enables more targeted strategic decisions, improving marketing mix effectiveness and enhancing customer experiences. The approach also addresses the scarcity and cost of expert labelers by utilizing generative AI for initial labeling and fine-tuning open-source models, thereby democratizing access to advanced analytical capabilities.

Beyond efficiency and effectiveness, synthetic specialists offer operational independence and reproducibility. Running efficiently on standard hardware, they reduce reliance on high-performance computing infrastructure and third-party providers. Their deterministic outputs secure reproducibility, addressing concerns about the non-deterministic nature of the more powerful generative AI models. This autonomy provides organizations with greater control over data and model behavior, preserving confidentiality and aligning with privacy regulations.

In sum, the introduction of synthetic specialists represents an advancement in harnessing AI for complex text classification tasks in marketing. By overcoming key limitations of large generative AI models, this approach facilitates the extraction of richer, more actionable insights from consumer-generated content. The study contributes valuable knowledge to both the academic community and industry practitioners, paving the way for more effective and sustainable applications of AI in marketing research and practice. Future research can build on this foundation, extending synthetic specialists to other domains and exploring their potential across other media types such as image, audio, and video.

REFERENCES

Abbasi, Ahmed, Jingjing Li, Donald Adjeroh, Marie Abate, and Wanhong Zheng (2019), "Don't mention it? Analyzing user-generated content signals for early adverse event warnings," *Information Systems Research*, 30 (3), 1007-28.

Aggarwal, Charu C, Alexander Hinneburg, and Daniel A Keim (2001), "On the surprising behavior of distance metrics in high dimensional space," in Database theory—ICDT 2001: 8th international conference London, UK, January 4–6, 2001 proceedings 8: Springer.

Archak, Nikolay, Anindya Ghose, and Panagiotis G Ipeirotis (2011), "Deriving the pricing power of product features by mining consumer reviews," *Management Science*, 57 (8), 1485-509.

Arora, Neeraj, Ishita Chakraborty, and Yohei Nishimura (2024), "EXPRESS: AI-Human Hybrids for Marketing Research: Leveraging LLMs as Collaborators," *Journal of Marketing*, 00222429241276529.

Baquero, Juan Aristi, Roger Burkhardt, Arvind Govindarajan, and Thomas Wallace (2020), "Derisking AI by design: How to build risk management into AI development," *McKinsey & Company*.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W Moe, Oded Netzer, and David A Schweidel (2020), "Uniting the tribes: Using text for marketing insight," *Journal of Marketing*, 84 (1), 1-25.

Brand, James, Ayelet Israeli, and Donald Ngwe (2023), "Using gpt for market research," *Available at SSRN 4395751*.

Busch, Kristen E. (2023), "Generative Artificial Intelligence and Data Privacy: A Primer," in Congressional Research Service: Congressional Research Service.

Camacho-Collados, Jose, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, and Eugenio Martínez-Cámara (2022), "TweetNLP: Cutting-edge natural language processing for social media," *arXiv preprint arXiv:2206.14774*.

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: Springer.

Castelo, Noah, Zsolt Katona, Peiyao Li, and Miklos Sarvary (2024), "How AI Outperforms Humans at Creative Idea Generation," *Available at SSRN 4751779*.

Chakraborty, Ishita, Minkyung Kim, and K Sudhir (2022), "Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes," *Journal of Marketing Research*, 59 (3), 600-22.

Chen, Xi, Quanquan Liu, and Yining Wang (2023), "Active Learning for Contextual Search with Binary Feedback," *Management Science*, 69 (4), 2165-81.

Chui, Michael, Roger Roberts, and Lareina Yee (2022), "Generative AI is here: How tools like ChatGPT could change your business," *Quantum Black AI by McKinsey*.

Cohn, David, Les Atlas, and Richard Ladner (1994), "Improving generalization with active learning," *Machine Learning*, 15, 201-21.

Daniels, Jodi (2023), "How Generative AI Can Affect Your Business' Data Privacy," in Forbes.

Desislavov, Radosvet, Fernando Martínez-Plumed, and José Hernández-Orallo (2023), "Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning," *Sustainable Computing: Informatics and Systems*, 38, 100857.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018), "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

Dew, Ryan, Asim Ansari, and Olivier Toubia (2022), "Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design," *Marketing Science*, 41 (2), 401-25.

Frankel, Richard, Jared Jennings, and Joshua Lee (2022), "Disclosure sentiment: Machine learning vs. dictionary methods," *Management Science*, 68 (7), 5514-32.

Goli, Ali and Amandeep Singh (2024), "Frontiers: Can Large Language Models Capture Human Preferences?," *Marketing Science*.

Guo, Biyang, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu (2023), "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," *arXiv preprint arXiv:2301.07597*.

Hamming, Richard W (1950), "Error detecting and error correcting codes," *The Bell system technical journal*, 29 (2), 147-60.

Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp (2023), "More than a feeling: Accuracy and application of sentiment analysis," *International Journal of Research in Marketing*, 40 (1), 75-87.

Hayes, Andrew F and Klaus Krippendorff (2007), "Answering the call for a standard reliability measure for coding data," *Communication Methods and Measures*, 1 (1), 77-89.

Horton, John J (2023), "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," *arXiv preprint arXiv:2301.07543*.

Hou, Wenpin and Zhicheng Ji (2024), "Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis," *Nature Methods*, 1-4.

Howard, Jeremy and Sebastian Ruder (2018), "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*.

Jürgensmeier, Lukas and Bernd Skiera (2024), "Generative AI for scalable feedback to multimodal exercises," *International journal of research in marketing*.

Kaufmann, Aviv (2024), "Understanding the Total Cost of Inferencing Large Language Models," Dell Technologies.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad (2014), "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research*, 50, 723-62.

Kotler, Philip, Bobby J Calder, Edward C Malthouse, and Peter J Korsten (2012), "The gap between the vision for marketing and reality," *MIT Sloan Management Review*, 53 (1), 13-14.

Kumar, Ajay and Tom Davenport (2023), "How to make generative AI greener," *Harvard Business Review*, 20.

Leffer, Lauren (2023), "When It Comes to AI Models, Bigger Isn't Always Better," in Scientific American. New York, NY: Springer Nature.

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.

Liaukonytė, Jūra, Anna Tuchman, and Xinrong Zhu (2023), "Frontiers: Spilling the beans on political consumerism: Do social media boycotts and buycotts translate to real sales impact?," *Marketing Science*, 42 (1), 11-25.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019), "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*.

Luangrath, Andrea Webb, 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.

Mallipeddi, Rakesh R, Subodha Kumar, Chelliah Sriskandarajah, and Yunxia Zhu (2022), "A framework for analyzing influencer marketing in social networks: selection and scheduling of influencers," *Management Science*, 68 (1), 75-104.

Manning, Christopher D (2022), "Human language understanding & reasoning," *Daedalus*, 151 (2), 127-38.

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, and Vanessa Parli (2023), "Artificial intelligence index report 2023," *arXiv preprint arXiv:2310.03715*.

McInnes, Leland, John Healy, and James Melville (2018), "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*.

Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen (2020), "BERTweet: A pre-trained language model for English Tweets," *arXiv preprint arXiv:2005.10200*.

Palatucci, Mark, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell (2009), "Zero-shot learning with semantic output codes," *Advances in neural information processing systems*, 22.

Pang, Bo and Lillian Lee (2008), "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval*, 2 (1–2), 1-135.

Peres, Renana, Martin Schreier, David Schweidel, and Alina Sorescu (2023), "On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice," *International Journal of Research in Marketing*, 40 (2), 269-75.

Puranam, Dinesh, Vrinda Kadiyali, and Vishal Narayan (2021), "The impact of increase in minimum wages on consumer perceptions of service: A transformer model of online restaurant reviews," *Marketing Science*, 40 (5), 985-1004.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018), "Improving language understanding by generative pre-training," in OpenAI Technical Report. San Francisco, CA: OpenAI Technical Report.

Reisenbichler, Martin, Thomas Reutterer, David A Schweidel, and Daniel Dan (2022), "Frontiers: Supporting content marketing with natural language generation," *Marketing Science*, 41 (3), 441-52.

Rocklage, Matthew D, Sharlene He, Derek D Rucker, and Loran F Nordgren (2023), "Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language," *Journal of Marketing Research*, 1, 19.

Rust, Roland T, Katherine N Lemon, and Valarie A Zeithaml (2004), "Return on marketing: Using customer equity to focus marketing strategy," *Journal of Marketing*, 68 (1), 109-27.

Rust, Roland T, William Rand, Ming-Hui Huang, Andrew T Stephen, Gillian Brooks, and Timur Chabuk (2021), "Real-time brand reputation tracking using social media," *Journal of Marketing*, 85 (4), 21-43.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl (2023), "Frontiers: Polarized America: From political polarization to preference polarization," *Marketing Science*, 42 (1), 48-60.

Şeref, Michelle MH, Onur Şeref, Alan S Abrahams, Shawndra B Hill, and Quinn Warnick (2023), "Rhetoric Mining: A New Text-Analytics Approach for Quantifying Persuasion," *INFORMS Journal on Data Science,*.

Shapiro, Benson P (1985), "Rejuvenating the marketing mix," *Harvard Business Review*, 63 (5), 28-34.

Snow, Rion, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng (2008), "Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks," in Proceedings of the 2008 conference on empirical methods in natural language processing.

Sundberg, Niklas (2024), "Tackling AI's Climate Change Problem," *MIT Sloan Management Review*, 65 (2), 38-41.

Suslava, Kate (2021), "'"Stiff business headwinds and uncharted economic waters": The use of euphemisms in earnings conference calls," *Management Science*, 67 (11), 7184-213.

van der Maaten, Laurens and Geoffrey Hinton (2008), "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 9 (Nov), 2579-605.

Van Dis, Eva AM, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting (2023), "ChatGPT: five priorities for research," *Nature*, 614 (7947), 224-26.

Van Noorden, R. and J. M. Perkel (2023), "AI and science: what 1,600 researchers think," *Nature*, 621 (7980), 672-75.

Van Waterschoot, Walter and Christophe Van den Bulte (1992), "The 4P classification of the marketing mix revisited," *Journal of Marketing*, 56 (4), 83-93.

Xia, Sibei and Chuanlan Liu (2022), "Applying Machine Learning to Study the Marketing Mix's Effectiveness in a Social Marketing Context: Fashion Brands' Twitter Activities in the Pandemic," *International Journal of Business Analytics (IJBAN)*, 9 (6), 1-17.

Appendix A – Calculations for Cost, Speed, and CO_2 Emissions

Supplier	meta	**Training H100 GPUs**	16,000	**Training electricity (MW)**	52,428
Model	llama-v3p1-405b-instruct	**Training GPU Hours**	30,840,000	**Training CO2 emissions (tons)**	19,346
Parameters	405 billion	**Training Days**	80	*corresponds to electricity for*	*14,979 households*

Inference Large: FP16

GPU	VRAM (GB)	GPU (count)	Watts /h	CO2 Kg/h	CapEx GPU invest	CapEx /h	Power /h	Own h	Rent /h	Output Tokens /sec
A6000 Ada 48GB	972	21	7,225	2.67	€ 84,956	2.70	€ 1.62	€ 4.32	€ 19.57	-
H100 SXM 80GB	972	13	7,225	2.67	€ 350,610	11.13	€ 1.62	€ 12.75	€ 43.79	-

Inference Quantized: INT4

	VRAM (GB)	GPU (count)	Watts /h	CO2 Kg/h	CapEx GPU invest	CapEx /h	Power /h	Own h	Rent /h	Output Tokens /sec
A6000 Ada 48GB	243	6	2,350	0.87	€ 24,273	0.77	€ 0.53	€ 1.30	€ 5.59	14
H100 SXM 80GB	243	4	2,500	0.92	€ 107,880	3.42	€ 0.56	€ 3.99	€ 13.47	28

Hosted: 3 models x 3 runs x 1000 Tweets (quantized INT4)

	Tweets /sec	Tweets /h	Time (minutes)	Cost own GPU	Cost rent GPU	kg CO2 Emission
A6000 Ada 48GB	0.11	400	150.00	€ 3.25	€ 13.98	€ 0.67
H100 SXM 80GB	0.16	560	107.14	€ 7.12	€ 24.06	€ 1.65
fixed cost for initial GPU purchase (own)		*A6000 48GB € 24,273*		*H100 80GB*	*€ 107,880*	

GenAI Pooling with 3 models x 3 runs x 1000 Tweets

	Tweets /sec	Tweets /h	Time* (minutes)	Cost API	kg CO2 Emission
Fireworks / OpenAI API	0.38	1383	43.40	€ 3.56	2.00

**concurrent API calls for models: llama 3.1 405B and 70B, GPT-4o, assuming 0.2 watt per internet query and, 50-watt laptop*

Synthetic Specialist: 1 x 1000 Tweets (fine-tuned RoBERTa 354m parameters)

	Tweets /sec	Tweets /h	Time (minutes)	Cost on MacBook	kg CO2 Emission
Local Laptop Computer	114.03	410,490	0.15	€ 0.00	0.0001
fixed cost for initial 15k genAI labels from 3 pooled genAIs		*650.93*	*€ 48.03*	*30.04*	

it takes approximately 22min to fine-tune RoBERTa with 15k Tweets at 100-watt on MacBook M2 Pro

Notes and Assumptions

369 g CO2 per kw/h usa mix (Wiatros-Motyka, M., N. Fulghum, and D. Jones, Global Electricity Review 2024, Ember, U.K.). 3,500 Kwh household consumption p.a. (BDEW Bundesverband der Energie- und Wasserwirtschaft e. V., 2024)
llama training 30.84M GPU hours (llama 3.1 model card on github)
llama training >16k GPUs (developer.nvidia.com/blog/supercharging-llama-3-1-across-nvidia-platforms/)
FP16: Number of parameters * 2 bytes * 1.2 | INT4 : Number of parameters * 0.5 bytes * 1.2
1.2 multiplier accounts for ~20% overhead for activations, gradients, and runtime
GPU count calculated by dividing the VRAM requirement by 80 (48) GB and rounding up
Power consumption: (Number of GPUs * Watts) + 400W + (Number of GPUs * 75W)
250W mean A6000 GPU inference consumption | 450W mean H100 GPU inference consumption
400W: Server base system consumption + 75W additional overhead per GPU for server components
GPUs are assumed to be running at a consistent load during inference. Cooling requirements are assumed to be included in these power estimates. The same number of GPUs is used for both FP16 and INT4 when the model fits on a single GPU.

GPU	CAPEX	llambda	runpod	CoreWeave	Mean
A6000 Ada 48GB	$4,500.00	$0.80	$1.03	$1.28	$1.04
H100 SXM 80GB	$30,000.00	$2.99	$3.49	$4.76	$3.75

Prices retrieved Sep 18, 2024
Assume 4 years lifetime per GPU
Synthetic Specialist: CAPEX 2,700USD macbook pro, 100 watt inference, 4 years lifetime
API usage assumes 0.2 watt per internet query with 50-watt laptop consumption
Electricity cost per 1000W: $0.25
USD to EUR: 0.899

Appendix B – Model Performance across Multiple Metrics

Common classification metrics in machine learning include Precision, Recall, and F1 Score. These can be calculated either across all 4Ps, or for individual P's, such as for Product, Place, Price, and Promotion. Precision measures the proportion of positive predictions that were correct, while Recall measures how many of the actual positives were accurately predicted. The F1 Score balances Precision and Recall, providing a single measure that reflects both.

Hamming Loss is used primarily for multilabel classification tasks. It measures the fraction of labels incorrectly predicted, accounting for both false positives and false negatives. A score of 0 indicates perfect predictions, while a score of 1 means every label was predicted incorrectly for each instance. Unlike Precision, Recall, and F1 Score, which focus on individual labels or instances, Hamming Loss considers all labels at once, providing a holistic view of a model's overall error rate.

Krippendorff's Alpha (α) measures the agreement between raters or models across all labels, accounting for the possibility of chance agreement. This metric captures the consistency of predictions across multiple labels and provides a robust measure of inter-rater or inter-model reliability.

We present all metrics for each model overall and by marketing mix element in Table B1.

Table B1. Agreement with Expert Labels by Marketing Mix Element

Model	MMX	F1 Score	Precision	Recall	H-loss	K-α
Amateurs mTurk	Overall	.695 *(.001)*	.762 *(.001)*	.638 *(.001)*	.194 *(.001)*	.552 *(.001)*
Amateurs mTurk	Product	.726 *(.001)*	.728 *(.001)*	.728 *(.001)*	.274 *(.001)*	.452 *(.003)*
Amateurs mTurk	Place	.750 *(.002)*	.808 *(.002)*	.727 *(.002)*	.181 *(.001)*	.500 *(.003)*
Amateurs mTurk	Price	.864 *(.002)*	.937 *(.001)*	.822 *(.002)*	.079 *(.001)*	.728 *(.003)*
Amateurs mTurk	Promotion	.733 *(.001)*	.761 *(.001)*	.727 *(.001)*	.244 *(.001)*	.467 *(.003)*
GenAI Pooling *(3 models)*	Overall	.876 *(.001)*	.844 *(.001)*	.910 *(.001)*	.090 *(.000)*	.806 *(.001)*
GenAI Pooling *(3 models)*	Product	.880 *(.001)*	.881 *(.001)*	.881 *(.001)*	.119 *(.001)*	.761 *(.002)*
GenAI Pooling *(3 models)*	Place	.910 *(.001)*	.907 *(.001)*	.913 *(.001)*	.075 *(.001)*	.820 *(.002)*
GenAI Pooling *(3 models)*	Price	.950 *(.001)*	.949 *(.001)*	.952 *(.001)*	.033 *(.001)*	.901 *(.002)*
GenAI Pooling *(3 models)*	Promotion	.867 *(.001)*	.866 *(.001)*	.878 *(.001)*	.131 *(.001)*	.734 *(.002)*
Phi-3-medium-4k-instruct	Overall	.681 *(.001)*	.622 *(.001)*	.753 *(.001)*	.245 *(.001)*	.483 *(.001)*
Phi-3-medium-4k-instruct	Product	.717 *(.001)*	.785 *(.001)*	.726 *(.001)*	.265 *(.001)*	.435 *(.003)*
Phi-3-medium-4k-instruct	Place	.795 *(.001)*	.789 *(.002)*	.803 *(.001)*	.173 *(.001)*	.590 *(.003)*
Phi-3-medium-4k-instruct	Price	.855 *(.001)*	.829 *(.001)*	.900 *(.001)*	.107 *(.001)*	.711 *(.002)*
Phi-3-medium-4k-instruct	Promotion	.543 *(.002)*	.704 *(.001)*	.628 *(.001)*	.434 *(.002)*	.087 *(.003)*
Synthetic Specialist	Overall	.896 *(.001)*	.877 *(.001)*	.915 *(.001)*	.074 *(.000)*	.839 *(.001)*

Synthetic Specialist	Product	.899 *(.001)*	.903 *(.001)*	.902 *(.001)*	.101 *(.001)*	.798 *(.002)*
Synthetic Specialist	Place	.926 *(.001)*	.935 *(.001)*	.919 *(.001)*	.059 *(.001)*	.853 *(.002)*
Synthetic Specialist	Price	.935 *(.001)*	.937 *(.001)*	.934 *(.001)*	.043 *(.001)*	.871 *(.002)*
Synthetic Specialist	Promotion	.905 *(.001)*	.903 *(.001)*	.908 *(.001)*	.092 *(.001)*	.810 *(.002)*
WizardLM-2-7B	Overall	.588 *(.001)*	.429 *(.001)*	.935 *(.001)*	.454 *(.001)*	.083 *(.002)*
WizardLM-2-7B	Product	.439 *(.001)*	.658 *(.002)*	.546 *(.001)*	.472 *(.001)*	-.122 *(.003)*
WizardLM-2-7B	Place	.484 *(.002)*	.632 *(.001)*	.616 *(.001)*	.515 *(.002)*	-.032 *(.003)*
WizardLM-2-7B	Price	.672 *(.002)*	.670 *(.001)*	.745 *(.002)*	.276 *(.001)*	.344 *(.003)*
WizardLM-2-7B	Promotion	.369 *(.001)*	.639 *(.003)*	.530 *(.001)*	.552 *(.002)*	-.261 *(.003)*
WizardLM-2-8x22B	Overall	.670 *(.001)*	.544 *(.001)*	.874 *(.001)*	.298 *(.001)*	.398 *(.001)*
WizardLM-2-8x22B	Product	.646 *(.002)*	.742 *(.001)*	.678 *(.001)*	.334 *(.002)*	.292 *(.003)*
WizardLM-2-8x22B	Place	.617 *(.001)*	.643 *(.001)*	.673 *(.002)*	.370 *(.001)*	.235 *(.003)*
WizardLM-2-8x22B	Price	.844 *(.001)*	.837 *(.002)*	.853 *(.001)*	.105 *(.001)*	.689 *(.003)*
WizardLM-2-8x22B	Promotion	.610 *(.002)*	.693 *(.001)*	.662 *(.001)*	.383 *(.002)*	.221 *(.003)*
claude-3-5-sonnet	Overall	.770 *(.001)*	.739 *(.001)*	.804 *(.001)*	.166 *(.001)*	.640 *(.001)*
claude-3-5-sonnet	Product	.822 *(.001)*	.834 *(.001)*	.821 *(.001)*	.175 *(.001)*	.645 *(.002)*
claude-3-5-sonnet	Place	.792 *(.001)*	.784 *(.001)*	.806 *(.001)*	.178 *(.001)*	.585 *(.003)*
claude-3-5-sonnet	Price	.907 *(.001)*	.937 *(.001)*	.883 *(.001)*	.058 *(.001)*	.814 *(.002)*
claude-3-5-sonnet	Promotion	.745 *(.001)*	.774 *(.001)*	.772 *(.001)*	.255 *(.001)*	.490 *(.003)*
claude-3-haiku	Overall	.660 *(.001)*	.535 *(.001)*	.861 *(.001)*	.307 *(.001)*	.380 *(.001)*
claude-3-haiku	Product	.494 *(.001)*	.671 *(.002)*	.574 *(.001)*	.442 *(.001)*	-.011 *(.003)*
claude-3-haiku	Place	.721 *(.002)*	.720 *(.002)*	.723 *(.002)*	.231 *(.001)*	.442 *(.003)*
claude-3-haiku	Price	.864 *(.001)*	.877 *(.001)*	.853 *(.001)*	.087 *(.001)*	.728 *(.003)*
claude-3-haiku	Promotion	.493 *(.002)*	.711 *(.001)*	.602 *(.001)*	.469 *(.002)*	-.013 *(.003)*
claude-3-opus	Overall	.721 *(.001)*	.631 *(.001)*	.842 *(.001)*	.225 *(.001)*	.532 *(.001)*
claude-3-opus	Product	.680 *(.001)*	.757 *(.001)*	.704 *(.001)*	.306 *(.001)*	.360 *(.003)*
claude-3-opus	Place	.757 *(.002)*	.817 *(.002)*	.733 *(.001)*	.176 *(.001)*	.514 *(.003)*
claude-3-opus	Price	.876 *(.001)*	.865 *(.001)*	.890 *(.001)*	.085 *(.001)*	.752 *(.002)*
claude-3-opus	Promotion	.661 *(.001)*	.744 *(.001)*	.710 *(.001)*	.334 *(.001)*	.323 *(.003)*
claude-3-sonnet	Overall	.678 *(.001)*	.591 *(.001)*	.794 *(.001)*	.262 *(.001)*	.458 *(.002)*
claude-3-sonnet	Product	.759 *(.001)*	.775 *(.001)*	.759 *(.001)*	.236 *(.001)*	.518 *(.003)*
claude-3-sonnet	Place	.583 *(.002)*	.630 *(.001)*	.653 *(.001)*	.411 *(.001)*	.166 *(.003)*
claude-3-sonnet	Price	.808 *(.001)*	.784 *(.001)*	.857 *(.001)*	.145 *(.001)*	.617 *(.003)*
claude-3-sonnet	Promotion	.744 *(.002)*	.775 *(.001)*	.773 *(.001)*	.255 *(.002)*	.489 *(.003)*
fine-tuned gpt-4o	Overall	.857 *(.001)*	.821 *(.001)*	.897 *(.001)*	.104 *(.000)*	.776 *(.001)*
fine-tuned gpt-4o	Product	.861 *(.001)*	.866 *(.001)*	.864 *(.001)*	.139 *(.001)*	.722 *(.002)*
fine-tuned gpt-4o	Place	.856 *(.001)*	.842 *(.001)*	.880 *(.001)*	.125 *(.001)*	.712 *(.002)*
fine-tuned gpt-4o	Price	.950 *(.001)*	.943 *(.001)*	.959 *(.001)*	.033 *(.001)*	.901 *(.002)*
fine-tuned gpt-4o	Promotion	.879 *(.001)*	.880 *(.001)*	.878 *(.001)*	.117 *(.001)*	.758 *(.002)*
fine-tuned gpt-4o mini	Overall	.820 *(.001)*	.805 *(.001)*	.835 *(.001)*	.127 *(.001)*	.721 *(.001)*
fine-tuned gpt-4o mini	Product	.806 *(.001)*	.823 *(.001)*	.812 *(.001)*	.193 *(.001)*	.612 *(.003)*
fine-tuned gpt-4o mini	Place	.860 *(.001)*	.850 *(.001)*	.873 *(.001)*	.119 *(.001)*	.720 *(.002)*
fine-tuned gpt-4o mini	Price	.942 *(.001)*	.949 *(.001)*	.936 *(.001)*	.038 *(.001)*	.885 *(.002)*
fine-tuned gpt-4o mini	Promotion	.828 *(.001)*	.856 *(.001)*	.818 *(.001)*	.159 *(.001)*	.656 *(.003)*
gemini-1.0-pro-002	Overall	.672 *(.001)*	.670 *(.001)*	.674 *(.001)*	.228 *(.001)*	.497 *(.001)*
gemini-1.0-pro-002	Product	.691 *(.002)*	.795 *(.001)*	.707 *(.001)*	.282 *(.001)*	.382 *(.003)*
gemini-1.0-pro-002	Place	.792 *(.001)*	.796 *(.001)*	.788 *(.001)*	.169 *(.001)*	.583 *(.003)*
gemini-1.0-pro-002	Price	.834 *(.002)*	.907 *(.001)*	.794 *(.002)*	.095 *(.001)*	.668 *(.003)*
gemini-1.0-pro-002	Promotion	.625 *(.002)*	.741 *(.001)*	.687 *(.001)*	.365 *(.002)*	.250 *(.003)*
gemini-1.5-flash-001	Overall	.718 *(.001)*	.670 *(.001)*	.773 *(.001)*	.210 *(.001)*	.550 *(.001)*
gemini-1.5-flash-001	Product	.770 *(.001)*	.815 *(.001)*	.772 *(.001)*	.220 *(.001)*	.540 *(.003)*
gemini-1.5-flash-001	Place	.812 *(.001)*	.807 *(.001)*	.818 *(.001)*	.157 *(.001)*	.624 *(.002)*
gemini-1.5-flash-001	Price	.879 *(.001)*	.897 *(.001)*	.864 *(.002)*	.077 *(.001)*	.757 *(.003)*
gemini-1.5-flash-001	Promotion	.597 *(.001)*	.750 *(.001)*	.671 *(.001)*	.387 *(.001)*	.193 *(.003)*
gemini-1.5-pro-001	Overall	.747 *(.001)*	.675 *(.001)*	.838 *(.001)*	.196 *(.001)*	.587 *(.001)*
gemini-1.5-pro-001	Product	.805 *(.001)*	.819 *(.001)*	.804 *(.001)*	.192 *(.001)*	.610 *(.002)*

gemini-1.5-pro-001	Place	.826 *(.001)*	.814 *(.001)*	.845 *(.001)*	.151 *(.001)*	.652 *(.002)*
gemini-1.5-pro-001	Price	.888 *(.001)*	.881 *(.001)*	.896 *(.001)*	.076 *(.001)*	.775 *(.003)*
gemini-1.5-pro-001	Promotion	.622 *(.002)*	.754 *(.001)*	.688 *(.001)*	.366 *(.002)*	.244 *(.003)*
gemma-2-27b-it	Overall	.773 *(.001)*	.723 *(.001)*	.831 *(.001)*	.169 *(.001)*	.638 *(.001)*
gemma-2-27b-it	Product	.810 *(.001)*	.812 *(.001)*	.812 *(.001)*	.190 *(.001)*	.620 *(.002)*
gemma-2-27b-it	Place	.814 *(.001)*	.820 *(.001)*	.809 *(.002)*	.151 *(.001)*	.628 *(.003)*
gemma-2-27b-it	Price	.887 *(.001)*	.897 *(.001)*	.878 *(.002)*	.073 *(.001)*	.775 *(.003)*
gemma-2-27b-it	Promotion	.737 *(.001)*	.766 *(.001)*	.764 *(.001)*	.263 *(.001)*	.474 *(.002)*
gemma-2-9b-it	Overall	.738 *(.001)*	.634 *(.001)*	.883 *(.001)*	.217 *(.001)*	.552 *(.001)*
gemma-2-9b-it	Product	.810 *(.001)*	.810 *(.001)*	.809 *(.001)*	.190 *(.001)*	.619 *(.002)*
gemma-2-9b-it	Place	.721 *(.001)*	.741 *(.001)*	.793 *(.001)*	.269 *(.001)*	.443 *(.002)*
gemma-2-9b-it	Price	.898 *(.001)*	.904 *(.001)*	.892 *(.001)*	.067 *(.001)*	.795 *(.003)*
gemma-2-9b-it	Promotion	.648 *(.002)*	.755 *(.001)*	.705 *(.001)*	.344 *(.002)*	.296 *(.003)*
gpt-3.5-turbo	Overall	.743 *(.001)*	.689 *(.001)*	.806 *(.001)*	.193 *(.001)*	.588 *(.001)*
gpt-3.5-turbo	Product	.792 *(.001)*	.808 *(.001)*	.792 *(.001)*	.204 *(.001)*	.584 *(.003)*
gpt-3.5-turbo	Place	.837 *(.001)*	.829 *(.001)*	.846 *(.001)*	.138 *(.001)*	.674 *(.002)*
gpt-3.5-turbo	Price	.869 *(.001)*	.867 *(.001)*	.872 *(.001)*	.087 *(.001)*	.739 *(.003)*
gpt-3.5-turbo	Promotion	.650 *(.002)*	.734 *(.001)*	.700 *(.001)*	.345 *(.002)*	.301 *(.003)*
gpt-4-turbo	Overall	.793 *(.001)*	.773 *(.001)*	.816 *(.001)*	.147 *(.001)*	.679 *(.001)*
gpt-4-turbo	Product	.817 *(.001)*	.822 *(.001)*	.816 *(.001)*	.182 *(.001)*	.634 *(.002)*
gpt-4-turbo	Place	.797 *(.001)*	.785 *(.001)*	.828 *(.001)*	.182 *(.001)*	.593 *(.003)*
gpt-4-turbo	Price	.913 *(.001)*	.916 *(.001)*	.911 *(.001)*	.057 *(.001)*	.826 *(.002)*
gpt-4-turbo	Promotion	.828 *(.001)*	.826 *(.001)*	.832 *(.001)*	.168 *(.001)*	.656 *(.002)*
gpt-4o	Overall	.855 *(.001)*	.834 *(.001)*	.878 *(.001)*	.103 *(.000)*	.775 *(.001)*
gpt-4o	Product	.857 *(.001)*	.857 *(.001)*	.858 *(.001)*	.142 *(.001)*	.715 *(.002)*
gpt-4o	Place	.875 *(.001)*	.867 *(.001)*	.887 *(.001)*	.105 *(.001)*	.751 *(.002)*
gpt-4o	Price	.954 *(.001)*	.950 *(.001)*	.959 *(.001)*	.031 *(.001)*	.908 *(.002)*
gpt-4o	Promotion	.863 *(.001)*	.861 *(.001)*	.868 *(.001)*	.134 *(.001)*	.727 *(.002)*
gpt-4o-mini	Overall	.783 *(.001)*	.717 *(.001)*	.863 *(.001)*	.166 *(.001)*	.649 *(.001)*
gpt-4o-mini	Product	.843 *(.001)*	.844 *(.001)*	.843 *(.001)*	.156 *(.001)*	.687 *(.002)*
gpt-4o-mini	Place	.865 *(.001)*	.870 *(.001)*	.861 *(.001)*	.110 *(.001)*	.730 *(.002)*
gpt-4o-mini	Price	.914 *(.001)*	.943 *(.001)*	.891 *(.002)*	.054 *(.001)*	.827 *(.003)*
gpt-4o-mini	Promotion	.648 *(.002)*	.764 *(.001)*	.708 *(.001)*	.343 *(.002)*	.297 *(.003)*
llama-v3p1-405b-instruct	Overall	.793 *(.001)*	.691 *(.001)*	.932 *(.001)*	.168 *(.001)*	.652 *(.001)*
llama-v3p1-405b-instruct	Product	.874 *(.001)*	.875 *(.001)*	.875 *(.001)*	.126 *(.001)*	.748 *(.002)*
llama-v3p1-405b-instruct	Place	.883 *(.001)*	.878 *(.001)*	.889 *(.001)*	.097 *(.001)*	.767 *(.002)*
llama-v3p1-405b-instruct	Price	.927 *(.001)*	.906 *(.001)*	.955 *(.001)*	.051 *(.001)*	.854 *(.002)*
llama-v3p1-405b-instruct	Promotion	.583 *(.002)*	.749 *(.001)*	.663 *(.001)*	.398 *(.002)*	.167 *(.003)*
llama-v3p1-70b-instruct	Overall	.808 *(.001)*	.835 *(.001)*	.783 *(.001)*	.129 *(.001)*	.711 *(.001)*
llama-v3p1-70b-instruct	Product	.826 *(.001)*	.827 *(.001)*	.827 *(.001)*	.174 *(.001)*	.652 *(.002)*
llama-v3p1-70b-instruct	Place	.871 *(.001)*	.879 *(.001)*	.865 *(.001)*	.104 *(.001)*	.743 *(.002)*
llama-v3p1-70b-instruct	Price	.889 *(.001)*	.928 *(.001)*	.862 *(.002)*	.068 *(.001)*	.779 *(.003)*
llama-v3p1-70b-instruct	Promotion	.819 *(.001)*	.835 *(.001)*	.812 *(.001)*	.169 *(.001)*	.639 *(.002)*
llama-v3p1-8b-instruct	Overall	.689 *(.001)*	.595 *(.001)*	.817 *(.001)*	.256 *(.001)*	.472 *(.001)*
llama-v3p1-8b-instruct	Product	.678 *(.002)*	.715 *(.002)*	.692 *(.001)*	.315 *(.001)*	.356 *(.003)*
llama-v3p1-8b-instruct	Place	.659 *(.001)*	.686 *(.001)*	.725 *(.001)*	.330 *(.001)*	.319 *(.003)*
llama-v3p1-8b-instruct	Price	.846 *(.001)*	.888 *(.001)*	.818 *(.002)*	.093 *(.001)*	.693 *(.003)*
llama-v3p1-8b-instruct	Promotion	.714 *(.001)*	.724 *(.001)*	.730 *(.001)*	.285 *(.001)*	.428 *(.003)*
mistral-large-2407	Overall	.748 *(.001)*	.623 *(.001)*	.934 *(.001)*	.219 *(.001)*	.555 *(.001)*
mistral-large-2407	Product	.683 *(.001)*	.762 *(.001)*	.707 *(.001)*	.303 *(.001)*	.366 *(.003)*
mistral-large-2407	Place	.874 *(.001)*	.868 *(.001)*	.881 *(.001)*	.106 *(.001)*	.747 *(.002)*
mistral-large-2407	Price	.914 *(.001)*	.902 *(.001)*	.927 *(.001)*	.059 *(.001)*	.828 *(.002)*
mistral-large-2407	Promotion	.572 *(.002)*	.749 *(.001)*	.656 *(.001)*	.406 *(.002)*	.145 *(.003)*
mistral-medium	Overall	.709 *(.001)*	.582 *(.001)*	.907 *(.001)*	.258 *(.001)*	.477 *(.001)*
mistral-medium	Product	.643 *(.001)*	.758 *(.001)*	.679 *(.001)*	.333 *(.001)*	.286 *(.003)*
mistral-medium	Place	.740 *(.001)*	.748 *(.001)*	.802 *(.001)*	.247 *(.001)*	.481 *(.003)*

mistral-medium	Price	.863 *(.001)*	.844 *(.002)*	.890 *(.001)*	.097 *(.001)*	.726 *(.003)*
mistral-medium	Promotion	.644 *(.001)*	.686 *(.001)*	.677 *(.001)*	.355 *(.001)*	.288 *(.003)*
o1-preview	Overall	.735 *(.001)*	.593 *(.001)*	.966 *(.001)*	.241 *(.001)*	.513 *(.001)*
o1-preview	Product	.644 *(.002)*	.762 *(.001)*	.680 *(.001)*	.332 *(.002)*	.287 *(.003)*
o1-preview	Place	.802 *(.001)*	.792 *(.001)*	.844 *(.001)*	.180 *(.001)*	.604 *(.003)*
o1-preview	Price	.909 *(.001)*	.880 *(.001)*	.954 *(.001)*	.066 *(.001)*	.818 *(.002)*
o1-preview	Promotion	.597 *(.002)*	.753 *(.001)*	.673 *(.001)*	.386 *(.002)*	.195 *(.003)*
open-mistral-nemo	Overall	.688 *(.001)*	.652 *(.001)*	.728 *(.001)*	.229 *(.001)*	.507 *(.001)*
open-mistral-nemo	Product	.718 *(.001)*	.748 *(.001)*	.722 *(.001)*	.272 *(.001)*	.437 *(.003)*
open-mistral-nemo	Place	.774 *(.001)*	.792 *(.002)*	.761 *(.001)*	.177 *(.001)*	.547 *(.003)*
open-mistral-nemo	Price	.859 *(.001)*	.860 *(.002)*	.858 *(.002)*	.093 *(.001)*	.718 *(.003)*
open-mistral-nemo	Promotion	.615 *(.002)*	.738 *(.001)*	.679 *(.001)*	.374 *(.002)*	.230 *(.003)*
open-mixtral-8x22b	Overall	.708 *(.001)*	.608 *(.001)*	.848 *(.001)*	.242 *(.001)*	.501 *(.001)*
open-mixtral-8x22b	Product	.563 *(.002)*	.736 *(.001)*	.624 *(.001)*	.391 *(.001)*	.126 *(.003)*
open-mixtral-8x22b	Place	.780 *(.001)*	.848 *(.001)*	.752 *(.001)*	.159 *(.001)*	.560 *(.003)*
open-mixtral-8x22b	Price	.883 *(.001)*	.888 *(.002)*	.879 *(.001)*	.076 *(.001)*	.767 *(.003)*
open-mixtral-8x22b	Promotion	.653 *(.001)*	.733 *(.001)*	.701 *(.001)*	.342 *(.001)*	.306 *(.003)*
open-mixtral-8x7b	Overall	.689 *(.001)*	.607 *(.001)*	.798 *(.001)*	.249 *(.001)*	.481 *(.002)*
open-mixtral-8x7b	Product	.680 *(.001)*	.748 *(.001)*	.702 *(.001)*	.308 *(.001)*	.360 *(.003)*
open-mixtral-8x7b	Place	.726 *(.001)*	.721 *(.001)*	.762 *(.001)*	.250 *(.001)*	.453 *(.003)*
open-mixtral-8x7b	Price	.838 *(.001)*	.853 *(.002)*	.825 *(.002)*	.103 *(.001)*	.675 *(.003)*
open-mixtral-8x7b	Promotion	.661 *(.002)*	.664 *(.002)*	.670 *(.002)*	.336 *(.002)*	.322 *(.003)*

Notes: N = 1000 Tweets (4000 labels); mean scores across 1,000 bootstraps with replacement; standard error in parentheses; K-α is Krippendorff's α; H-loss is Hamming Loss.