Academic Research Project:

Evaluating the Performance of Large Language Models in Marketing

Maria Fernanda Rodriguez Tamez [redacted]

Department of Marketing, Strategy & Entrepreneurship, University Canada West

MBAR 661: Academic Research Project (ONS-SPRING25-04)

Mohsen Ghodrat

May 30th 2025

Table of Contents

Abstract

As Large Language Models (LLMs) become more integrated into marketing, evaluating their performance in context-specific scenarios is essential. This study examines four leading LLMs—ChatGPT-4, Claude 3, Gemini 2.5 Pro, and LLaMA 3.1—on 50 short-form marketing tasks. Using a dual evaluation framework, we compare model outputs through both LLM-as-judge and human-as-judge methods, scoring performance on clarity, relevance, creativity, and persuasive impact. While LLMs often generate fluent, human-like responses, they show varying success in emotional tone and brand alignment, especially in sensitive contexts. Human-written responses remained stronger in empathy and nuance. This research offers practical insights for marketers and emphasizes the importance of human oversight when using LLMs in emotionally resonant content.

Keywords: LLMs, marketing communication, generative AI, emotional tone, content quality.

Introduction

The rise of Large Language Models (LLMs) in marketing has opened a new chapter in content creation, customer interaction, and campaign design. Though their main advantage lies in speed and scalability, the most critical element in any marketing effort remains the emotional and psychological connection with consumers. Marketing is not merely the transfer of information—it's a strategic effort to earn trust, inspire emotion, and influence how a brand is perceived.

These models, which are trained on massive datasets and can produce fluent and sensible text, are now being applied to tasks such as writing product descriptions and ad copy, personalizing email campaigns, and even helping in customer service automation. Their ability to produce human-like responses, draw insights from behavioral data, and operate at scale makes them attractive tools for marketing professionals under increasing pressure to generate more content, more quickly, and for more channels than ever before.

However, marketing poses different challenges for LLMs than the type of utility tasks typically used to assess Natural Language Processing (NLP) models. While the general NLP benchmarks (building on MMLU or BIG-Bench) are about factual accuracy or grammatical precision, marketing requires tone sensitivity, emotional resonance, creativity, and alignment with brand values. In this context, a technically correct sentence may still fail if it lacks persuasive power, misaligns with brand identity, or comes across as emotionally tone-deaf. Evaluating LLMs in marketing, therefore, must consider more than output fluency—it must assess strategic effectiveness.

Recent developments have moved LLMs closer to human-level performance in a broad range of text-based tasks, yet concerns linger about their trustworthiness and suitability for marketing contexts where emotive stakes and culturally sensitive messaging are high. A hallucination could risk appearance inexperience, inconsistent tone or lack of context sensitivity could actually cost consumer trust and brand perception. On the other end, the cost-efficiency and consistency of LLM-generated content offer clear advantages, especially when human teams are constrained by time or resources.

We assess how effectively four top-performing LLMs—ChatGPT-4, Claude 3, Gemini 2.5 Pro, and LLaMA 3.1—respond to 50 realistic marketing tasks, focusing on their performance and adaptability. Based on academic marketing theory and AI evaluation literature, we propose a dual evaluation approach that pairs machine-driven judgments and human-driven judgments. Our approach rates model outputs in terms of four key dimensions: clarity, relevance, creativity, and persuasiveness. In the process, we examine not only which responses perform best, but also how closely machine-generated marketing language approximates that of a human expert.

In this context, assessing LLMs in marketing is more than simply monitoring their technical fluency—it's about how these systems align with the brand tone, emotional nuance, and the communicative strategy behind business goals. This study proposes a two-pathway evaluation framework that combines rank-ordering and human-likeness assessments to evaluate model performance across 50 marketing scenarios. By using both LLM-based and human-based evaluation strategies, the research explores how well state-of-the-art models can replicate the tone, intent, and quality of professional marketing writing.

While preliminary in scope, this study offers useful observations on how LLMs perform in key marketing scenarios, highlighting areas where human input remains critical and where automated tools may provide support.

Literature Review

Large language models (LLMs) have rapidly evolved from research innovations to revolutionary tools that are transforming and shaping industries today. Marketing is one industry that is especially affected since LLM's capacity to produce writing that is human-like, personalized content and draw conclusions from a lot of data from customers' insights provides previously unseen benefits. LLMs are becoming a key component in marketing tasks from creating ad material and customizing emails to evaluating market research and producing fake consumer reactions.

However, the use of LLMs in marketing is not simple. Brand tone consistency, factual accuracy, emotional appeal, ethical sensitivity, and measurable business impact are all unique requirements for marketing. These needs exceed the conventional criteria for natural language processing (such as BLEU or ROUGE) by a significant amount. Therefore, it is necessary to tailor the assessment of LLMs in marketing to take into consideration both, technical capabilities and real-world marketing objectives.

History and Evolution of Large Language Models

The development of chatbot systems started decades ago with early rule-based systems such as (1966), which was a straightforward but significant chatbot that mimicked psychotherapist discussions using predefined templates (Bommasani et al., 2021). Despite being revolutionary, ELIZA was not able to comprehend the language. Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks were two significant developments in the 1990s that improved the processing of sequential text input (Bommasani et al., 2021). Yet, training deep language networks was still hard because gradients disappeared and there wasn't enough data.

The Transformer architecture ("Attention is All You Need") was introduced in 2017, marking an important milestone in natural language processing (Vaswani et al., 2017).

Transformers replaced recurrence with self-attention, allowing models to handle many parts of a text at once and making it possible to train on very large collections of text. Soon after, major discoveries began to emerge. In 2018, BERT launched masked language modeling, which greatly enhanced robots's comprehension of phrase context (Devlin et al., 2019).

Around the same time, the first version of the GPT series emerged, demonstrating that autoregressive transformers could produce unexpectedly solid and fluid language at scale. This progress continued with GPT-2, which had 1.5 billion parameters, and GPT-3, which expanded to 175 billion parameters (Bommasani et al., 2021). Millions of users gained access to these capabilities through the launch of ChatGPT in late 2022, based on GPT-3.5.

Next came GPT-4 in 2023, which introduced multimodal characteristics (accepting both text and pictures) and achieved results higher than human standards in a variety of academic and professional tasks (OpenAI, 2023). Other significant competitors joined the market alongside these, providing more open-source alternatives and alternative designs, such as LLaMA, Anthropic's Claude, and Meta's OPT-175B. As the research moves forward toward hybrid, multimodal, and agent-based systems, language production is only one part of a larger toolbox that also includes tool usage, reasoning, and adapting to an interactive environment.

Training, Model Scale, and Adaptation

Modern LLMs are remarkable in their scope. Over the past five years, models have grown from a few million parameters to hundreds of billions, with some experimental architectures now approaching the trillion-parameter mark (Brown 2020; Chowdhery 2022; Fedus, 2022). In this context, parameters refer to the internal weights that a neural network adjusts during training to learn patterns in language. These values are what allow the model to generalize from training data to new inputs. Scaling is based on the straightforward principle that a model with more parameters has a greater capacity to learn nuanced details, retain complex structures, and perform well across a variety of tasks (Zhang et al., 2022).

However, scale alone is insufficient. Multi-stage training paradigms are also crucial. Pre-training is the initial stage in which the model uses a self-supervised goal, usually next-token prediction, to process large text datasets. Without explicit task instructions, models develop a general comprehension of language and facts at this phase (Brown et al., 2020; Zhang et al., 2022).

After that, the model is fine-tuned by being further trained on carefully chosen datasets to specialize its behavior. For example, fine-tuning might educate a model to write in a certain brand tone, summarize news stories, or answer questions. For some usage scenarios, this greatly increases the model's dependability (Ouyang et al., 2022).

A key approach is instruction tuning, in which models are taught using prompts and desired results to improve their ability to follow human directions. Early users of this method, such as InstructGPT, significantly increased user satisfaction (Ouyang et al., 2022).

Another crucial stage is Reinforcement Learning from Human Feedback (RLHF), especially for chat-based models. In this case, outputs are ranked by human evaluators, and the model is further refined through reinforcement learning to favor answers that are more consistent with human values (Ouyang et al., 2022; Bai et al., 2022).

Furthermore, multimodal models like GPT-4 and Flamingo (DeepMind), which take pictures in addition to text as input, represent the most recent frontier. These models are a first step toward more comprehensive AI systems that are capable of reasoning with text, graphics, video, and even code (OpenAI, 2023).

Prompt engineering, few-shot learning, and retrieval-augmented generation (RAG) are strategies researchers use to make LLMs more practical. Prompt engineering guides the model's reactions. Few-shot learning provides examples in the prompt instead of retraining. RAG allows the model to use outside databases or knowledge sources to anchor responses in actual facts and reduce hallucinations (Lewis et al., 2020).

Current Issues and Biases in LLMs

Despite their incredible advancements, LLMs are still far from perfect. Hallucination, in which the model generates text that appears convincing but is actually inaccurate, is one of the most well-known problems (Lewis et al., 2020). LLMs may "make up" information to keep the conversation going since they are trained to guess the next most probable word rather than to check the accuracy of the material.

Advantages of LLMs

LLMs have many advantages. One of their greatest is the ability to generate human-like content at scale. Whether it’s email content, product description, ad copy, or blog posts, LLMs can effortlessly and naturally generate such content (Aghaei et al., 2024). They are great at personalization as well. LLMs may enhance relevance and engagement by personalizing messages to specific segmentation groups in accordance with an analysis of consumer data (Pearson, 2024). They are also used in chatbots and virtual assistants to provide iterations that represent always current customer support (Spajić et al., 2023).

LLMs are time-saving and cost-effective because they minimize the time and cost of brainstorming, writing, and/or condensing content. They are multilingual and can communicate with audiences around the world. Their power to turn unstructured data into actionable insights for strategy and planning is arguably its most impressive application (Pearson, 2024).

Limitations of LLMs

Overreliance on LLMs is a notable limitation. When human oversight is absent, the output may quickly become formulaic, misaligned with brand identity, or emotionally tone-deaf. Without careful review, there is a risk of losing the emotional nuance, sensitivity, and creativity essential for effective marketing communication.

Data privacy and intellectual property are also growing issues. Given that LLMs are trained on very large amounts of data, much of it wide-ranging and ambiguous data, the origin of what was "learned" can be difficult to trace or verify (German, 2024).

Performance Evaluation in Marketing

Even if writing produced by a model is technically correct, it might not suit the audience, match the brand tone, or cause offense. This is why evaluation must account for business outcomes and communication objectives (Aghaei, 2024).

LLMs can produce emotionally tone-deaf material, reinforce existing prejudice, or create real hallucinations (Spajić, 2023). Evaluation is also needed to compare candidate models and choose between options like open-source (LLaMA) and proprietary (GPT-4, Claude) (Federiakin, 2024). Knowing where the models do well and where they struggle, like with handling slang or sarcasm, helps reduce potential risks.

Types of evaluation include automatic benchmarks such as BLEU and ROUGE, which are commonly used in natural language processing to measure the overlap between generated and reference texts. BLEU (Bilingual Evaluation Understudy) evaluates precision by comparing n-gram matches, while ROUGE (Recall-Oriented Understudy for Gisting Evaluation) focuses more on recall, particularly for summarization tasks (Papineni et al., 2002; Lin, 2004). These metrics offer quantitative insights but often overlook subtleties like emotional tone, contextual appropriateness, or brand alignment—elements essential in marketing content. Therefore, human judgment remains crucial to assess qualities that automated scores may miss. Additional methods include prompt fidelity checks, sentiment and emotional coherence ratings, and key business metrics such as engagement, conversions, and retention.

Psychometric methods like Item Response Theory (IRT) offer more refined insights (Federiakin, 2024). LLM leaderboards like Hugging Face collapse performance into a single number (Federiakin, 2024). More robust systems can show the best models for tasks like customer service or content creation.

Proposed Framework

During this initial phase, the research focuses on identifying the most suitable framework for evaluating the performance of Large Language Models (LLMs) in marketing-related applications. Two promising paths are drawn from academic and strategic marketing literature. Both offer different advantages depending on the analytical scope

The first approach, the 7 Ps of Marketing, emerged from foundational marketing theory first introduced by Booms and Bitner (1981) It remains one of the most recognized and widely used models in academic and professional literature, which includes seven components: product, price, place, promotion, people, process, and physical evidence. The model is also ideal for a complete overview of the full marketing cycle, from product production to post-purchase feedback. Each factor can be mapped to tasks that LLMs are increasingly used to perform.

For instance, in the promotion category, LLMs could be assessed based on their ability to develop ad copy, social media captions, and customized email campaigns. Within the product category, models can help create product descriptions and brand positioning. Their responsibilities in the price-and-people-focused areas may involve writing persuasive pricing text or writing customer service scripts.

The use of the 7 Ps model facilitates the comprehensive evaluation of the LLM's efficacy in performing an interconnected set of marketing activities. This structure enhances both the coherence and practical relevance of the evaluation, especially considering the model's familiarity among both scholars and industry professionals. Therefore, it serves as an effective organizing principle for mapping LLM applications within real-world marketing scenarios.

The second approach was introduced during the early ideation phase of the project through exploratory interactions with ChatGPT. This model focuses on assessing fundamental marketing functions such as market research, segmentation, and consumer behavior analysis, or broader strategic objectives such as brand awareness, customer retention, or engagement. The functional- or objective-based model allows for direct alignment between LLM outputs and performance measures such as those that researchers might use to evaluate how content generated by a model can contribute directly to business results. This model would be particularly useful when the research is focused on examining the measurable impact of LLMs on marketing effectiveness, campaign ROI, or customer experience.

While the 7 Ps framework provides the main structure for organizing and evaluating the marketing tasks performed by LLMs, this study also draws on the function- and objective-based model to add depth to the evaluation. The 7 Ps are useful for sorting tasks, writing ad copy in promotion, and writing customer service messages to people. Yet the means-ends model, the function objective model, can also add value by enabling us to determine what success would look like for each project. It shifts the focus toward outcomes, such as whether the content improves engagement, communicates the brand message clearly, or helps attract new customers.

By integrating these two approaches, the study benefits from both structural clarity and alignment with practical marketing objectives, resulting in a more complete and contextually relevant evaluation of LLMs in marketing.

Methodology

For the purpose of this review, the LLMs were evaluated using the structural-outcome combined approach proposed in this study, and the methodological steps undertaken in the review were as follows. To investigate the effectiveness of large language models (LLMs) in performing practical marketing tasks, we adopted a multi-pathway and mixed-methods approach. This approach seeks to value model creativity and alignment with marketing goals while balancing automatic measurement with human judgment. It is based on two complementary strategies: an original LLM-as-judge model selection method, and a human expert-as-judge comparative evaluation approach.

1. Research Design

As discussed in the prior sections, LLMs show great potential in marketing communication tasks. However, their performance cannot be fully evaluated using traditional

Natural Language Processing (NLP) benchmarks. In order to evaluate the efficacy of LLMs in typical real-world marketing tasks, we developed a customized benchmarking framework based on a combination of academic evaluation methodologies and common business objectives.

Standard NLP benchmarks, e.g. MMLU (Hendrycks et al., 2021), HELM (Liang et al., 2022), and BIG-Bench (Srivastava et al., 2022) typically focus on objective tasks, including answering factual questions or performing logical and numerical calculations. These benchmarks rely on output-to-ground-truth comparisons which can be scalable and reproducible. However, such approaches are inadequate for what matters most in marketing, which is the effectiveness of communication, emotional impact, creativity, and proper integration into brand strategy.

The same applies to marketing content such as promotional emails, ad copy, or customer service scripts, which must be strategically crafted with a bit of creativity. These tasks do not have a single correct answer; instead, responses are judged based on how well they achieve a particular goal, such as capturing attention or responding to a complaint with an appropriate tone. As a result, conventional benchmarking methods are inadequate for evaluating the quality and strategic effectiveness of LLM-generated marketing content.

2. Evaluation Framework

To address this gap, we propose a multi-pathway evaluation framework that draws from both language model assessment methods and marketing literature. It includes several evaluation strategies, including scoring against a standardized rubric (covering clarity, relevance, creativity, and persuasive impact), as well as a "human-likeness" assessment, where judges determine whether a response was produced by a human or an AI.

Each prompt includes a human-generated response to serve as an anchor for comparison. In the original evaluation approach, ChatGPT-4 was also used as an impartial judge to blind-evaluate all responses, providing a simulated machine-to-machine evaluation layer. This automated judgment step reflects real-world use cases where LLMs not only generate marketing content but are also capable of ranking or refining outputs internally. In contrast, the refined human-as-judge approach places emphasis on subjective evaluation by expert marketers to assess the strategic and creative quality of each response.

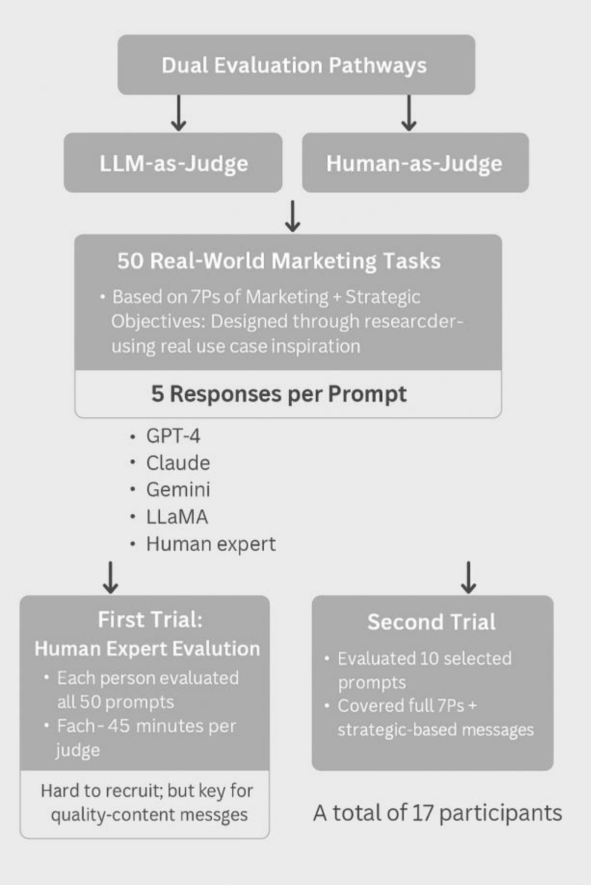

Dual Evaluation Pathways
LLM-as-Judge
Human-as-Judge
50 Real-World Marketing Tasks
• Based on 7Ps of Marketing + Strategic Objectives: Designed through researcder-using real use case inspiration
5 Responses per Prompt
• GPT-4
• Claude
• Gemini
• LLaMA
• Human expert
First Trial:
Human Expert Evalution
• Each person evaluated all 50 prompts
• Fach- 45 minutes per judge
Hard to recruit; but key for quality-content messges
Second Trial
• Evaluated 10 selected prompts
• Covered full 7Ps + strategic-based messages
A total of 17 participants

Figure 1. Diagram illustration of the dual evaluation methodology used in this study, showing the LLM-as-Judge and Human-as-Judge pathways applied to 50 marketing prompts with 5 responses each.

LLM-as-Judge Method

In the initial version of the LLM-as-judge strategy, the questions and multiple-choice answers (A–E) were produced by ChatGPT to be linguistically fluent and challenging. Yet each question and its set of possible answers were carefully reviewed, fine-tuned, and approved by the researcher before release. This enabled the model to be closely aligned with marketing reality in both theoretical terms and practical reality. The questions were then finalized and given to the four LLMs and the human expert, with a request that each would now choose the "best" answer. Here, the human expert judged but did not provide open-ended responses.

Human-as-Judge Method

In the refined human-as-judge approach, the role of the human expert shifted. Here, the expert wrote their original responses to the prompts that were used to generate LLMs' responses. The answers were added to a new multi-choice set and were reviewed by five independent new human judges. Within this setup, the human-generated responses served as a qualitative anchor —a benchmark for professional tone, relevance, and creativity against which to evaluate machine-generated alternatives.

The inclusion of a human expert in both steps added depth and realism to the study. Their involvement assisted us in exploring the gray areas of marketing communication – the subtle tone change, emotional persuasion, cultural responsiveness, and strategic contextualizing – that LLMs sometimes find challenging to mimic. In a time when LLMs are increasingly front-and-center in content generation, this human comparison was crucial for pointing out where these models excel — and where human intuition still leads.

3. Model Selection

This study compares the performance of four top-performing Large Language Models (LLMs) and one human marketing expert. Specifically, we selected ChatGPT-4, Claude 3, Gemini 2.5 Pro, and LLaMA 3.1 Nemotron 70B, representing a diverse range of scalable architectures with varying degrees of industry relevance and popularity.

ChatGPT-4 was included due to its consistent top-tier performance in industry benchmarks and its strong instruction-following behavior. It is also the last ranger in terms of evaluation due to the extensive tuning with Reinforcement Learning from Human Feedback (RLHF), which enables it to predict human preferences in content and evaluation scenarios (Ouyang, 2022; OpenAI, 2023).

Claude 3, developed by Anthropic, is known for its ethical sensitivity and its capacity for handling long-context inputs—features especially important in brand communication and customer experience scenarios. Gemini 2.5 Pro comes to a substantiation with facts for precise and unambiguous communication. LLaMA 3.1 Nemotron 70B, an open-source model, was added to provide transparency and serve as a counterbalance to proprietary systems. While it is less specialized, it contributes a valuable perspective from the non-commercial side of LLM development (Federiakin, 2024; Pearson, 2024).

4. Questions design

Due to time and resource limitations, the initial strategy of incorporating 100 open-ended questions was modified to the final number of 50 multiple-choice questions. These questions were designed around realistic marketing scenarios that involved the 7 Ps of Marketing and three broad marketing objectives. This structure offered a more scalable and evaluable approach while retaining the practical complexity needed to challenge both LLMs and human evaluators.

All the questions were developed by the researcher and then refined using ChatGPT to improve language and wording, without contributing to the core content or ideas. Each prompt was less than 100 words and also included five response options (A-E), all 25 to 30 words in length, to approximate real-world examples like ad copy, product blurbs, or the subject of an email. The answers were intentionally crafted to be similar in tone, quality, and structure, making the task of choosing the best option more demanding and realistic.

To ensure clarity, a standardized instruction prompt was developed and shared with each participant to minimize ambiguity and guide expectations. The prompts were sent in clusters of five to avoid token overload and streamline the response process. The prompts were then forwarded individually to the four LLMs and the human expert for completion. Each participant received only the scenario and instructions—none had access to the responses generated by others.

5. Evaluation Process

This study implemented two evaluation approaches to assess the quality of LLM-generated marketing content: one using a language model (ChatGPT-4) as the judge, and the other using a panel of human marketing professionals.

In the LLM-as-Judge strategy, ChatGPT-4 was assigned the role of evaluator. Thanks to its alignment with human preferences through reinforcement learning (Ouyang, 2022; OpenAI, 2023), it was selected to review anonymized responses and choose the best answer for each question. Each response was assessed based on predefined criteria: clarity, tone, persuasiveness, and strategic alignment. This method simulates real-world workflows in which LLMs are not only used to generate content but also to evaluate and refine it internally.

In the Human-as-Judge strategy, five independent marketing professionals were asked to evaluate the same anonymized response sets. Prompts were carefully reviewed to ensure they were neutral and task-relevant, while all system-generated outputs—including those from the human expert—were presented in random order. Judges were asked to select the most “appealing” answer for each marketing scenario, based on their own judgment and experience.

The concise length of the answers demanded careful attention from the human evaluators, who often found it challenging to choose a clear "best" response. In many cases, judges reported being unable to distinguish which answer had been written by the human expert, illustrating how advanced LLMs have become in generating content that closely mimics professional quality.

Together, these evaluation processes ensured fairness, replicability, and academic rigor, and at the same time captured the uncertainty and the subtlety that define real-world marketing communication.

Results

Discussion

LLM-as-Judge Results

In this section, we report key patterns from the original evaluation procedure, where each LLM ranked the best response from a list of five options per prompt. Though this approach did not include rubric-based scoring or extensive written justifications of human judges, we can still extract meaningful patterns by examining agreement levels between LLMs and the human expert.

On the 50 evaluated prompts, the four LLMs chose the same answer in 14% of the cases. When including the human expert, full consensus—where all LLMs and the human agreed—only occurred in 4% of the prompts. This highlights the diversity in how models and human experts interpret what constitutes the "best" response in a marketing scenario.

Interestingly, Gemini 2.5 Pro was the model that most frequently matched the human's selections, aligning with the expert in 22 out of 50 prompts (44%). This may suggest that Gemini is slightly more attuned to the human perspective, possibly due to its fact-based, instruction-sensitive behavior. Claude and ChatGPT followed closely behind, matching the human expert in 20 (40%) and 19 (38%) prompts respectively, while LLaMA showed the lowest alignment, matching in only 16 (32%).

Task-Specific Agreement Trends

In an examination of category-based patterns of agreement, the highest level of concordance between LLMs and the human expert was for "Product" and "Customer Support." For instance, the highest degree of agreement was found amongst questions 2, 3, and 4, which were about writing product descriptions and value propositions: for these questions, all four LLMs matched the human at least three times. This may be because these tasks are more objective and descriptive in nature, relying on clear product features and benefits.

Categories such as "Promotion" and "Tone and Brand Voice," however, saw significantly less agreement. For these questions, LLMs and the human expert frequently diverged, perhaps because these questions involve more emotional nuance, creativity, and persuasive framing, where machine outputs still struggle to fully capture human intuition.

Response Patterns: Human vs. LLM

One notable insight is that in many prompts, the human-written response did not stand out distinctly. In 68% of prompts, at least one LLM selected the human's answer as the “best,” despite not knowing its origin. This suggests that current-generation LLMs can produce content that is often indistinguishable in style or perceived quality from that of a human expert—especially in short-form tasks like product blurbs or support messages.

Human-as-Judge Results

The outcome in the human-as-judge evaluation provides an interesting conclusion on how perception on marketing content goes beyond technical correctness. Although the assessment for clarity, relevance, creativity, and persuasive effect, it was clear that the most successful answers were those that were not just cognitive, but also emotional.

Emotional Resonance and Consumer-Centric Communication

For some of the prompts, especially the ones related to health, lifestyle, identity-driven brands, etc (Q2: Vegan Skincare or Q10: Instagram Sale Caption), participants almost always chose the responses with a more emotive tone or narrative. These responses frequently featured evocative language, aspirational messaging , and brand-consistent voice that helped to create a sense of connection. Take, for example, the emotionally charged phrases “unapologetically bold” or “your glow has no borders” which not only describe a product but invite the reader to associate with a lifestyle. That underscores a crucial marketing rule: when you create an emotional connection, you create something memorable and engaging — especially in businesses like beauty, fashion, and wellness.

Interestingly, participants did not consistently identify the human-written responses. In several instances (e.g., Q1, Q4), the human data interspersed smoothly with model outputs. What this points to, then, is a new state of affairs in which LLMs are effectively equal in tone, structure, and fluency for many short-form marketing tasks. However, for prompts related to empathy, apology, or gratitude (e.g., Q15—Q17), the human voice was still more recognizable. When responses were written by people, they were warmer, more elusive, more erratic — especially when the aim was to soothe or patch up with customers. LLMs find it difficult to fully express these distinctions, especially in emotive contexts.

Clarity and Strategic Simplicity

The top-performing responses were not always the most stylistically complex, but instead the most confident and straightforward. In very simple prompts—"Q8 (Holiday Subjectline)" or Q20 (Refund Confirmation)" participants strongly favored a response that were actionable, concise, and clear. These decisions align with digital copywriting best practices in which short and sweet tends to beat out verbosity and message clarity is a direct correlation with the ability to convert.

Even in advertising, strategic simplicity had power. Lots of participants liked options that clearly conveyed the offer upfront, or that had an element of urgency: "Run, don't scroll 🏃 Everything is 30% off, no exceptions." Such statements weren't just information — the lines were delivered with a certain tone and speed aimed at influencing consumer behavior. The fact they were often model-generated indicates that LLMs are learning more and more about functional marketing patterns — not necessarily, however, about the emotional undercurrents that lead to consumer trust.

Human-Likeness and the Blurred Line

One surprising result was that participants on the whole found it quite hard to determine which response was the one created by a human. It's a reflection of how LLMs are increasingly mimicking the cadence, gloss , and intentionality of professional marketers. But it also raises the question of what, in the end, makes human writing special. If and when they guessed wrong or were surprised, they may have pointed to a tightening for its own sake, but possibly also for overvaluing polish as a signal of quality.

In reality, what made the human-written responses stand out—when they did—was not grammar or vocabulary, but tone sensitivity. Emotional calibration, cultural resonance, and intuitive timing remain challenging for LLMs. They may generate polished language, but contextual emotional relevance often requires more than just lexical fluency—it demands human instinct.

Implications for Marketing Practice

The results reconfirm that LLMs are indeed very strong content generators—yet only in a certain spectrum of tone and task. They are more effective when the goal of communication is clear, the structure is conventional, and the emotional stakes are not high. But when it comes to brand storytelling, identity-forming, and recovering the customer message, human input is still essential.

These insights suggest a future where LLMs serve as strategic partners rather than replacements. The brands that will succeed will be those that use LLMs to scale and standardize, using human oversight to ensure depth, emotional nuance, and brand authenticity.

Reflections on the Ethical Dimension of LLM Use in Marketing

LLMs are becoming more than just tools. They are increasingly integrated within the model of consumer interaction with companies, the way they make decisions, and the way they experience the product or service they consume. Such LLMs not only sell a product or service, they also mold perception, manipulate self-image, and often guide emotional choices.

When customers are being hit from all sides all the time, this is more important than ever. Consumers are continually influenced, pressured, or even carried away—by the messages they see and hear every day. This places a significant responsibility on those shaping those narratives. The evaluation of LLMs in advertising must be grounded in honesty and ethical responsibility. It's not just about performance, it's about building trust, ensuring transparency, and protecting the long-term relationship between people and technology.

Limitations of the Study

While this study offers valuable insights, there are some limitations to this research. First, the evaluation relied on a single human marketing expert to provide benchmark responses, which limits generalizability across different writing styles, brand voices, or industry contexts. Including multiple human experts in future iterations could better capture variation in professional judgment.

Second, the sample size of human judges used in the Human-as-Judge phase was limited, and only a subset of the scoring criteria (e.g., rubric-based dimensions) were rated by all participants. This leads to some variation in the perception and prevents a high level of quantification.

Third, all evaluations focused on short-form marketing content. Real-world use cases for this are numerous, but how well the framework adapts to longer forms of text (e.g. blog posts, ad scripts, product guides) remains unknown.

Finally, the models studied were up to date in 2025; however, the rapid development in the LLM field redirects these outcomes to change promptly, with the appearance of new developments, and their adoption.

Conclusions

Key Insights

This research validates that LLMs are rapidly closing the gap with human marketers in producing fluent, relevant, and sometimes emotionally resonant content. However, human judgment still plays a critical role, particularly in areas requiring subtle tone, empathy, or contextual sensitivity.

Framework Contribution

The evaluation framework introduced in this study bridges technical benchmarks with real-world communication goals. The framework offers a multi-dimensional lens for evaluating language quality, strategic fit, and audience perception. The inclusion of human-authored responses as qualitative anchors enhanced the depth of comparison, allowing evaluators to assess nuance that traditional metrics often overlook.

Practical Relevance

The findings offer some practical advice for marketing teams considering the integration of AI. LLMs are quite good at creating content in high-velocity, transactional use cases (subject lines, CTAs, product blurbs), but certainly require some human input in emotionally laden, brand-defining use cases like apologies, gratitude messages, or wellness storytelling. As such, businesses should see LLMs not as replacements but as powerful collaborators—ones that require strategic steering to maintain authenticity and emotional alignment.

Future Exploration

Future research could extend this framework to long-form content, multilingual outputs, or multimodal campaigns that combine text with visuals. Incorporating affective computing techniques or psycholinguistic tagging may also help LLMs improve emotional calibration. Moreover, deeper exploration of prompt engineering variables—such as temperature settings, persona modeling, or few-shot chains—could offer valuable insights into how to guide tone and intent in more controlled and brand-specific ways.

References

Aghaei, R., Tannous, K., & Renz, A. (2024). AI for marketing: Bridging creativity and data. *Journal of Digital Marketing, 15*(2), 133–149.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Olsson, C. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

Federiakin, D. (2024). Evaluating LLM performance for practical applications: From benchmarks to business impact. *AI and Society Review, 12*(1), 25–47.

German, D. (2024). Data provenance and accountability in generative models. *Ethics in AI Quarterly, 6*(1), 55–70.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Hashimoto, T. (2022). Holistic evaluation of language models.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Guu, K., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks.

OpenAI. (2023). *GPT-4 technical report*. https://openai.com/research/gpt-4

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Christiano, P. (2022). Training language models to follow instructions with human feedback.

Pearson, M. (2024). LLMs in customer experience and brand engagement. *Journal of AI in Business Strategy, 3*(1), 73–90.

Spajić, M., Zarić, T., & Radosavljević, M. (2023). Ethical challenges of generative AI in marketing: A case for responsible automation. *Journal of Business Ethics and Technology, 10*(3), 112–129.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*

Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text summarization branches out: Proceedings of the ACL-04 workshop*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... & Ott, M. (2022). OPT: Open pre-trained transformer language models.

Boom, B. H., & Bitner, M. J. (1981). Marketing strategies and organizational structures for service firms. In J. H. Donnelly & W. R. George (Eds.), *Marketing of services* (pp. 47–51). American Marketing Association.

Appendix

Appendix A: Complete list of 50 Marketing Scenario Prompts

The following marketing prompts were used to generate responses from four LLMs and one human expert. Each prompt simulates a real-world marketing task and requires a short, focused response of approximately 25–30 words.

PRODUCT *Subcategory: Product Descriptions and Differentiation*

Q1. Write a 25–30 word product description for an ergonomic chair designed for remote workers who sit for long hours. Focus on comfort, posture support, and modern aesthetics.

Q2. Craft a short product description for a vegan skincare brand that emphasizes both ethical values and luxury appeal.

Q3. Describe a unique selling proposition (USP) for a pair of noise-canceling headphones that automatically adjust to ambient noise levels.

Q4. Create an email launch teaser (25–30 words) for a smart water bottle that tracks hydration and glows to remind users to drink.

Q5. Write a short comparison between a basic and a premium smartwatch model.

PRICE *Subcategory: Promotional Messaging and Discounts*

Q6. Write a homepage banner message for a flash sale.

Q7. Create a headline for a 2-for-1 fitness gear promo.

Q8. Suggest a subject line for a 20% off holiday sale.

Q9. Write an email CTA for a 25% skincare sale.

Q10. Write an Instagram caption promoting a 30% storewide sale.

PLACE *Subcategory: Store and Delivery Information*

Q11. Write a friendly message for a store locator page inviting users to visit in person.

Q12. Write a short furniture delivery message focused on speed and convenience.

Q13. Create a message encouraging users to choose local pickup.

Q14. Announce that your company now ships internationally.

Q15. Write a message explaining a shipping delay due to high order volume.

PEOPLE *Subcategory: Customer Service Scripts*

Q16. Respond to a customer who wants to return a gently used item.

Q17. Apologize for an order that is five days late.

Q18. Thank a loyal customer for their kind words.

Q19. Apologize and offer a solution for sending the wrong item.

Q20. Confirm that a refund has been processed.

PROCESS

Subcategory: FAQ and Support Content

Q21. Answer a FAQ about possible shipping delays.

Q22. Provide step-by-step return instructions in FAQ format.

Q23. Troubleshoot login issues for a customer.

Q24. Write a short message confirming cancellation of a subscription.

Q25. Update a customer on the status of a refund.

PHYSICAL EVIDENCE

Subcategory: Digital Brand Experience and Touchpoints Q26. Describe a premium fashion e-commerce website's brand tone and user experience.

Q27. List trust elements that should appear at checkout.

Q28. Describe key elements of a homepage that reflects strong brand identity.

Q29. Write a welcome message for first-time site visitors.

Q30. Describe support portal features that make finding help easy.

PROMOTION

Subcategory: Email Campaigns, CTAs, and Engagement

Q31. Suggest a subject line to re-engage a lapsed customer.

Q32. Write a follow-up message to thank a customer for a repeat purchase.

Q33. Write a subscription cancellation confirmation message with a friendly tone.

Q34. Write a tagline for a wellness brand.

Q35. Write a mission statement for a sustainable lifestyle brand.

Q36. Describe how to keep a consistent brand voice across web and social media.

MARKET RESEARCH & CONSUMER INSIGHT

Subcategory: Persona Development and Insights

Q37. Share one insight about how customers perceive sustainability claims.

Q38. Create a persona profile for a health-conscious consumer.

Q39. Identify behaviors to build a segment for a nighttime wellness campaign.

FINAL INTEGRATIVE / STRATEGIC FIT

Evaluation Focus: Alignment with brand tone, message clarity, persuasiveness, and overall marketing coherence.

Q40. Write a short promotional message for a new sleep supplement.

Q41. Craft a CTA for a 20% off summer sale in an email header.

Q42. Craft a Call to Action (CTA) that creates urgency in an email offering 20% off.

Q43. Write a compelling subject line that opens a re-engagement campaign.

Q44. Suggest a structure for a promotional email that includes a time-limited offer.

Q45. Write a message to encourage in-store pickup over shipping.

Q46. Create messaging to promote a local store-exclusive event.

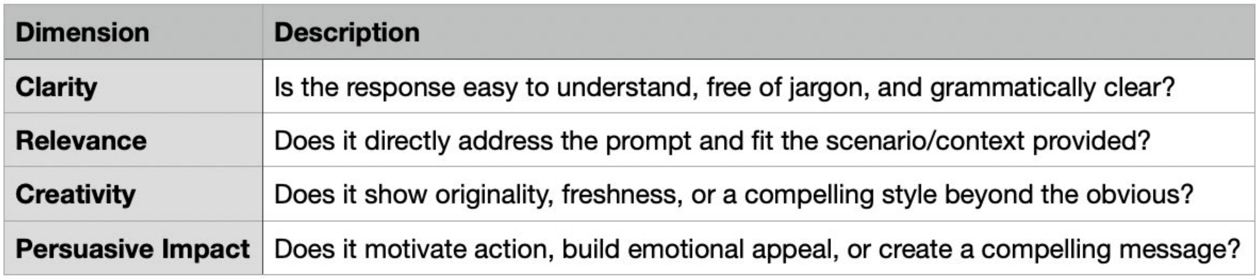
Q47. Suggest packaging copy that reflects strong eco-values.

Q48. Write a persuasive Call to Action (CTA) for a referral program or campaign.

Q49. Craft an Instagram CTA to boost post engagement through comments.

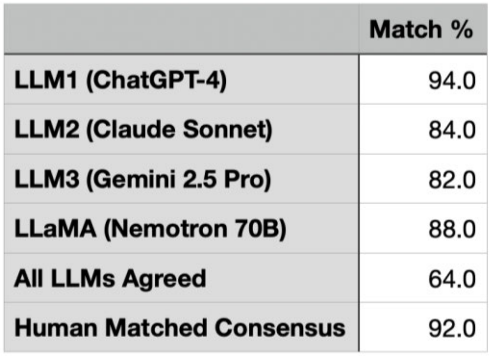
Q50. Write a short message aligning a new fragrance with a bold brand tone.

Appendix B - Evaluation Rubric



Dimension	Description
Clarity	Is the response easy to understand, free of jargon, and grammatically clear?
Relevance	Does it directly address the prompt and fit the scenario/context provided?
Creativity	Does it show originality, freshness, or a compelling style beyond the obvious?
Persuasive Impact	Does it motivate action, build emotional appeal, or create a compelling message?

Appendix C - LLM Human Alignment Summary



	Match %
LLM1 (ChatGPT-4)	94.0
LLM2 (Claude Sonnet)	84.0
LLM3 (Gemini 2.5 Pro)	82.0
LLaMA (Nemotron 70B)	88.0
All LLMs Agreed	64.0
Human Matched Consensus	92.0

Appendix D - Prompting Template Used for LLMs and Human Expert

The following instruction was provided to all participants (four LLMs and the human expert) to ensure consistency in task framing and expected output:

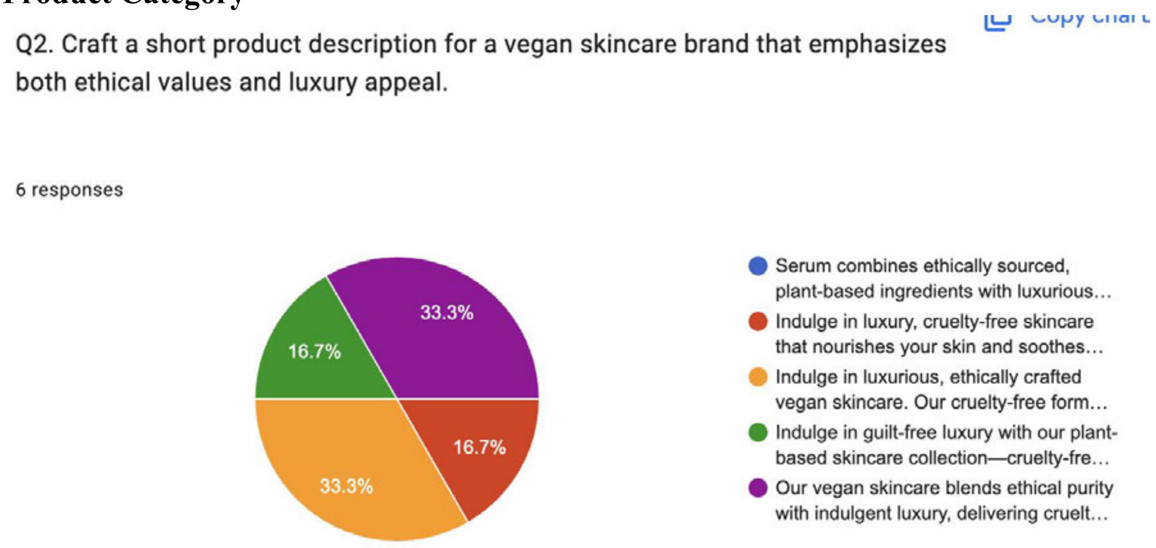
Instruction Prompt:

> “You are a professional copywriter with expertise in marketing communication. You will be given a series of short marketing scenarios. For each one, write a **25–30-word response** that fits the described goal. Responses should be concise, engaging, and aligned with industry best practices. Avoid redundancy and ensure your answer is appropriate for a general consumer audience.”

Each prompt was sent in clusters of five to avoid token overload and ensure smoother processing for the LLMs.

Appendix E - Comparison

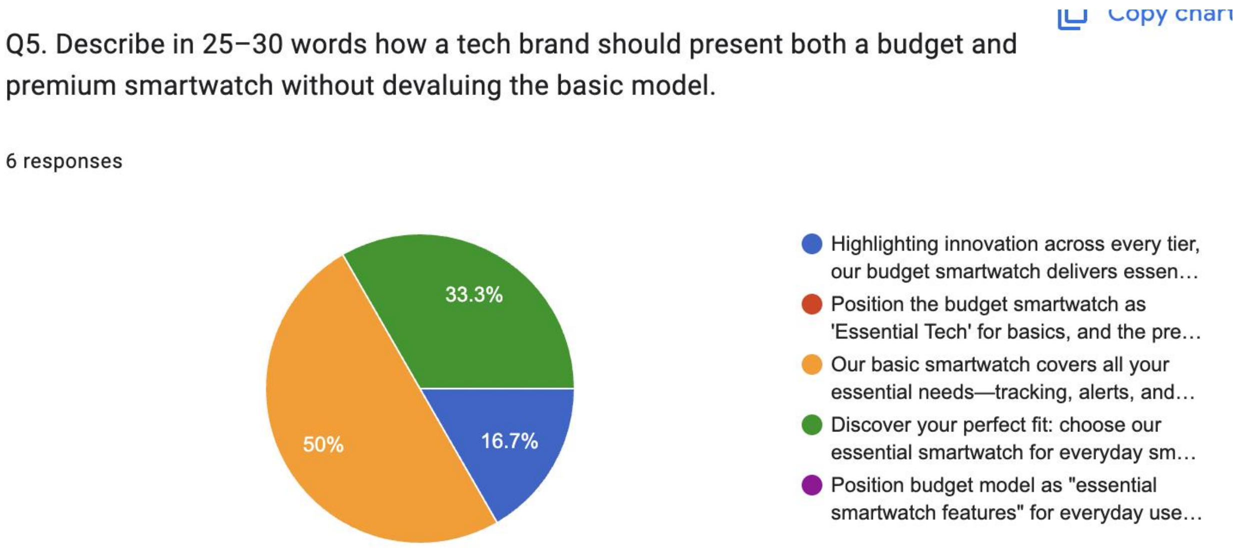
Product Category



Q2. Craft a short product description for a vegan skincare brand that emphasizes both ethical values and luxury appeal.

6 responses

Price Category

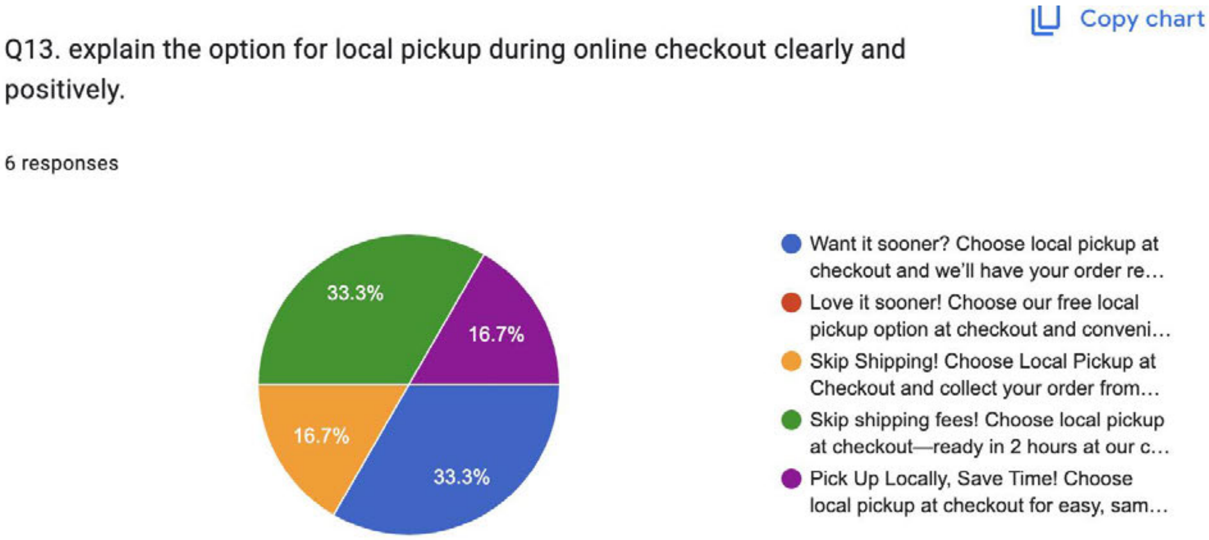


Q5. Describe in 25–30 words how a tech brand should present both a budget and premium smartwatch without devaluing the basic model.

6 responses

- ● Highlighting innovation across every tier, our budget smartwatch delivers essen…
- ● Position the budget smartwatch as 'Essential Tech' for basics, and the pre…
- ● Our basic smartwatch covers all your essential needs—tracking, alerts, and…
- ● Discover your perfect fit: choose our essential smartwatch for everyday sm…
- ● Position budget model as "essential smartwatch features" for everyday use…

Place Category

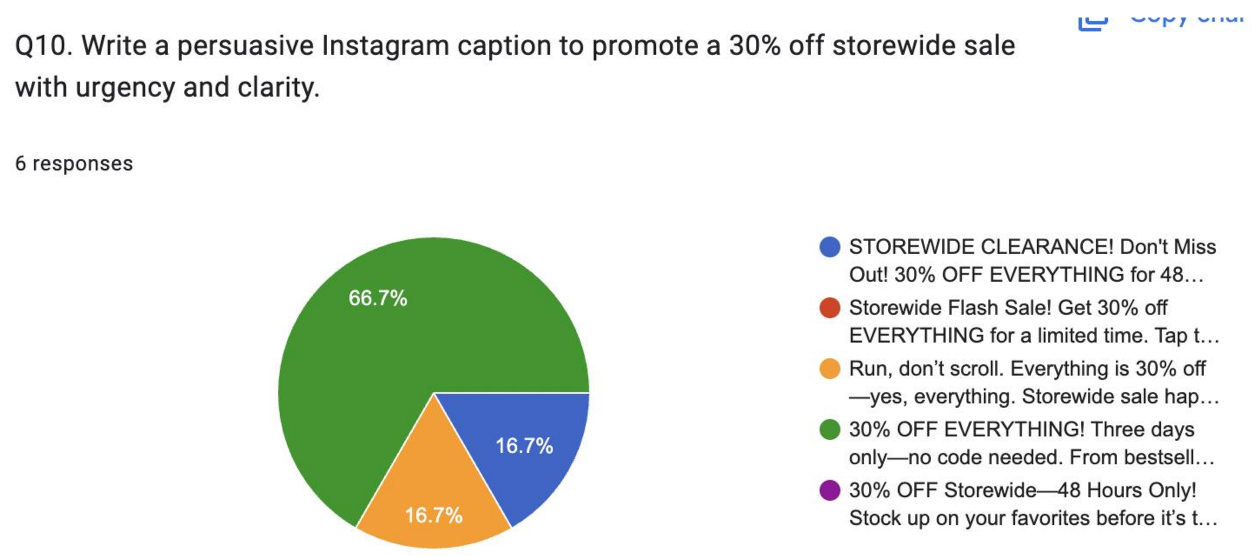


Q13. explain the option for local pickup during online checkout clearly and positively.

6 responses

- ● Want it sooner? Choose local pickup at checkout and we'll have your order re…
- ● Love it sooner! Choose our free local pickup option at checkout and conveni…
- ● Skip Shipping! Choose Local Pickup at Checkout and collect your order from…
- ● Skip shipping fees! Choose local pickup at checkout—ready in 2 hours at our c…
- ● Pick Up Locally, Save Time! Choose local pickup at checkout for easy, sam…

Promotion Category



Q10. Write a persuasive Instagram caption to promote a 30% off storewide sale with urgency and clarity.

6 responses

● STOREWIDE CLEARANCE! Don't Miss Out! 30% OFF EVERYTHING for 48…

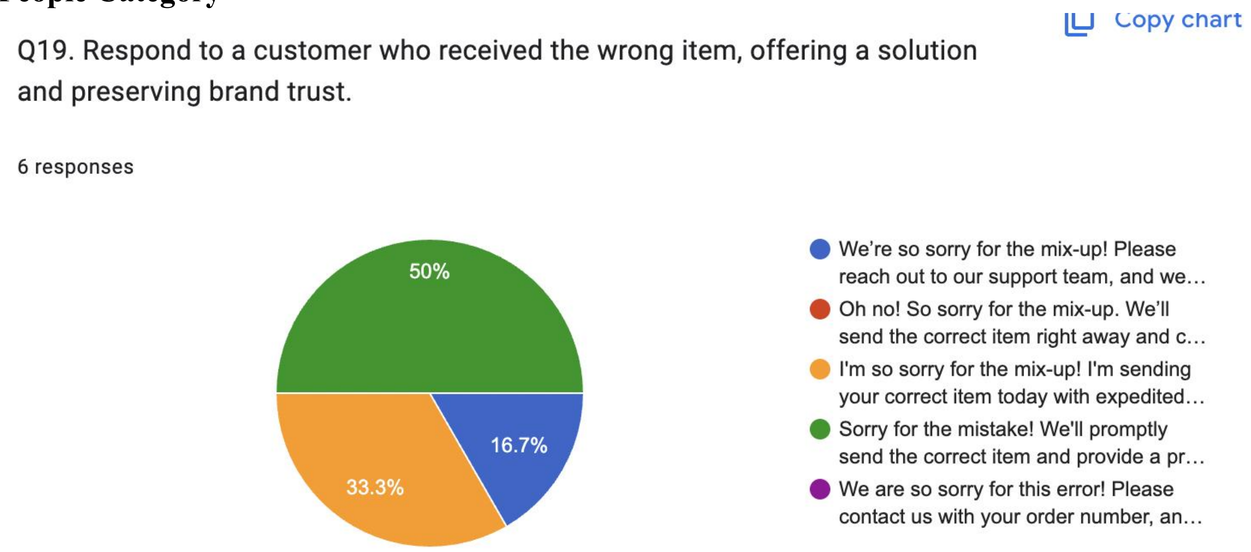
● Storewide Flash Sale! Get 30% off EVERYTHING for a limited time. Tap t…

● Run, don’t scroll. Everything is 30% off —yes, everything. Storewide sale hap…

● 30% OFF EVERYTHING! Three days only—no code needed. From bestsell…

● 30% OFF Storewide—48 Hours Only! Stock up on your favorites before it’s t…

People Category



Q19. Respond to a customer who received the wrong item, offering a solution and preserving brand trust.

6 responses

● We’re so sorry for the mix-up! Please reach out to our support team, and we…

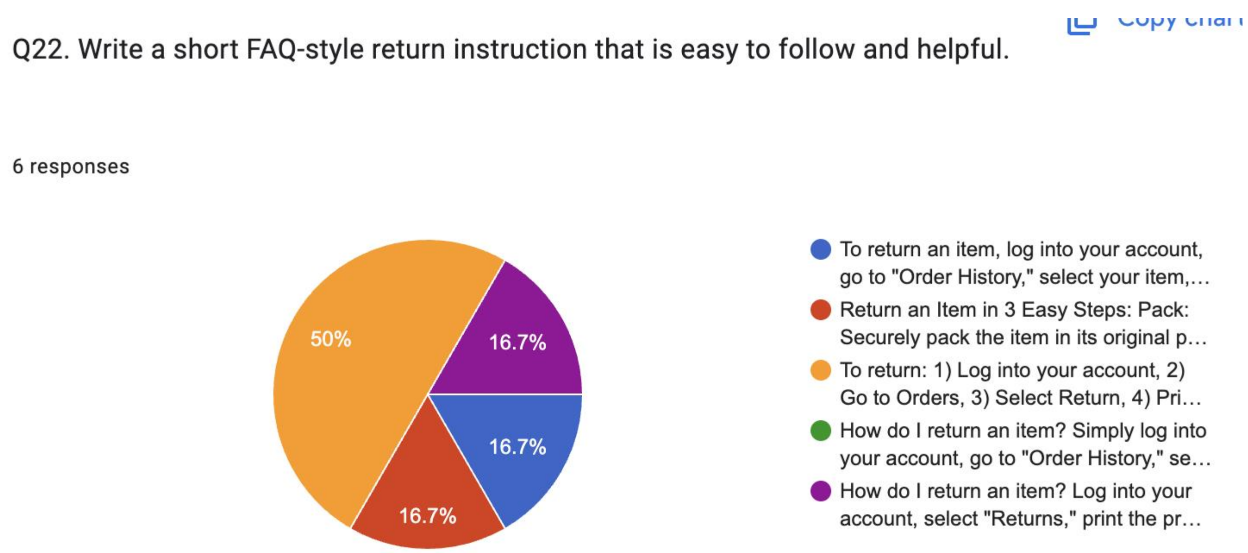
● Oh no! So sorry for the mix-up. We’ll send the correct item right away and c…

● I'm so sorry for the mix-up! I'm sending your correct item today with expedited…

● Sorry for the mistake! We'll promptly send the correct item and provide a pr…

● We are so sorry for this error! Please contact us with your order number, an…

Process Category

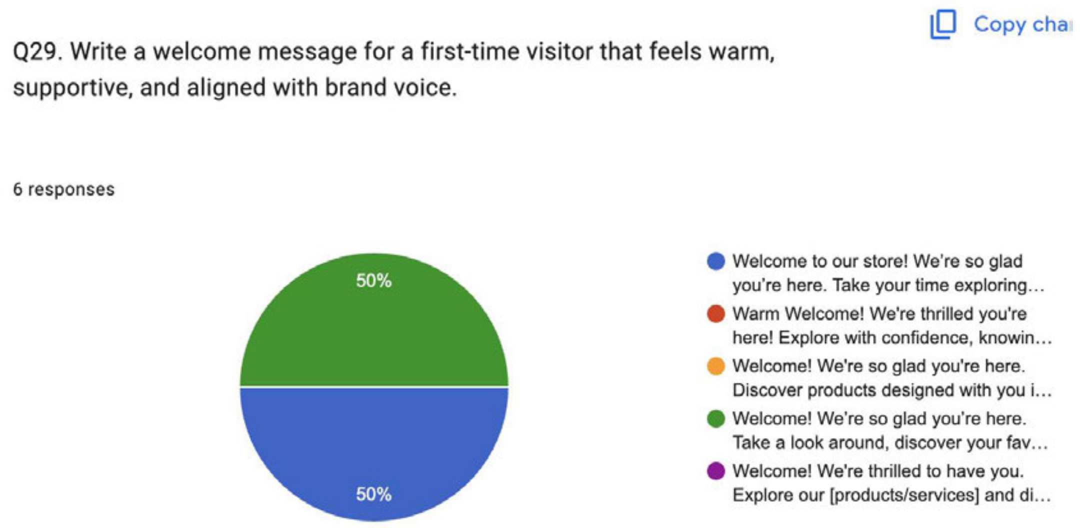


Q22. Write a short FAQ-style return instruction that is easy to follow and helpful.

6 responses

Copy chart
50%
16.7%
16.7%
16.7%
● To return an item, log into your account, go to "Order History," select your item,…
● Return an Item in 3 Easy Steps: Pack: Securely pack the item in its original p…
● To return: 1) Log into your account, 2) Go to Orders, 3) Select Return, 4) Pri…
● How do I return an item? Simply log into your account, go to "Order History," se…
● How do I return an item? Log into your account, select "Returns," print the pr…

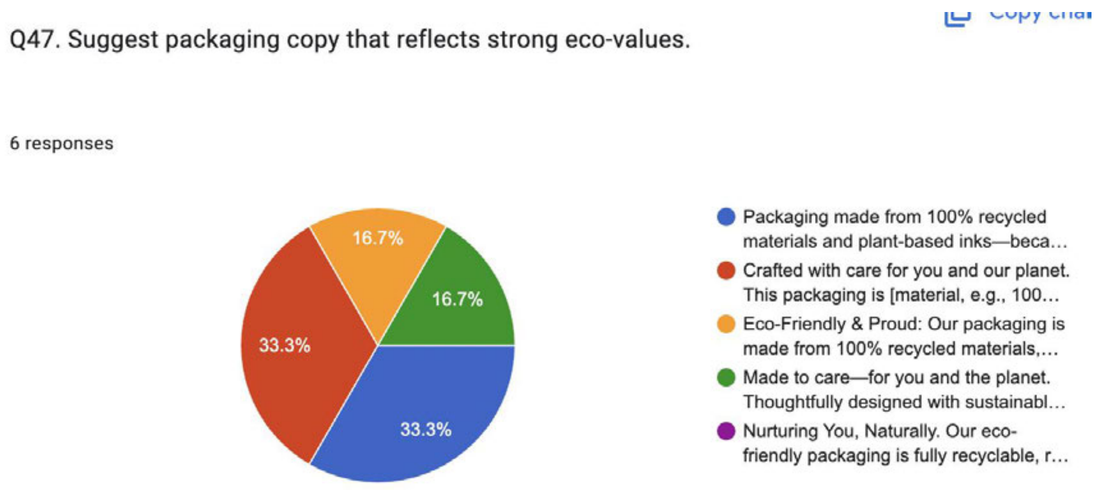
Comparison



Q29. Write a welcome message for a first-time visitor that feels warm, supportive, and aligned with brand voice.

6 responses

In Q29, which asked for a warm, brand-aligned welcome message, the human expert's response (blue) received 50% of the votes, while the remaining half went to a very similar LLM-generated response. This split suggests that while human-written content still resonates strongly in emotionally driven scenarios, some LLMs are now producing messaging nearly indistinguishable from human tone and intent. The fact that other model outputs were largely dismissed also shows that consistency across LLMs in capturing emotional nuance still varies.



Q47. Suggest packaging copy that reflects strong eco-values.

6 responses

While, this prompt resulted in a highly fragmented distribution of votes, with participants choosing different responses almost evenly. Unlike Q29, there was no dominant winner, and no clear indication of which response felt the most human-like or effective. This divergence illustrates that LLMs are capable of generating multiple compelling outputs, each resonating with different evaluators for different reasons. Rather than revealing a weakness, this pattern may actually highlight the strength of LLMs in producing diverse, high-quality messaging in areas where interpretation is subjective, such as sustainability values.