Issues in Information Systems
Volume 27, Issue 4, pp. 224-237 2026

DOI: https://doi.org/10.48009/4_iis_2026_119

Linguistic and emotional characteristics of LLM-generated marketing texts: A comparative analysis using Empath

Lisa Kovalchick, *Pennsylvania Western University, kovalchick@pennwest.edu*
Alan Peslak, *Penn State University, arp14@psu.edu*

Abstract

Large language models (LLMs) such as GPT-5.1, Gemini Flash 2.5, and Perplexity Sonar are increasingly used to generate marketing content. This study sets out to compare linguistic and emotional characteristics of LLM-generated marketing texts by analyzing 90 marketing texts generated from 30 standardized prompts submitted to the three aforementioned LLMs under deterministic conditions. Using Empath, an open-source lexical analysis tool, we extracted 194 semantic categories and identified 15 variables representing emotion, persuasion, achievement, and business-relevant dimensions. Prior to inferential testing, floor-effect checks identified three categories (negative emotion, joy, and fear) with insufficient variance for analysis. Across the remaining 12 testable categories, MANOVA indicated no significant multivariate differences. Univariate ANOVA with Bonferroni and Benjamini–Hochberg corrections revealed no statistically significant differences after correction for multiple comparisons. A full exploratory scan of all 194 Empath categories confirmed this null result. ANCOVA controlling significant word count differences across models further weakened observed trends. Cross-prompt stability analysis revealed that, in this dataset, Gemini produced the most consistent linguistic style across prompt types, while Perplexity exhibited the most variability. Thus, although the models converge in their average linguistic profiles, they remain distinguishable in the stability of those profiles resulting in divergence in variability rather than in central tendency. These findings suggest that, in this sample, contemporary LLMs produce broadly convergent marketing language under deterministic prompting conditions, with implications for model selection, prompt engineering, and automated marketing communication. Due to the nature of the study our results should only be interpreted as exploratory research using a preliminary, researcher-developed prompt corpus.

Keywords: large language models, marketing communication, Empath, psycholinguistics, LLM comparison, natural language generation

Introduction

"Traditional mass marketing strategies have become inadequate in meeting the ever-changing demands of modern consumers, who expect personalized content and interactions" (Büyüksomer, et al., 2024, p. 109). In the past, such personalized marketing strategies were unfathomable, due to the pure number of man hours required. However, in recent years, with the emergence and widespread availability of artificial intelligence (AI), personalized marketing is becoming an everyday occurrence.

Large Language Models (LLMs), a specialized type of AI that are designed to comprehend and generate human-like text based on the prompts they are given (Stanford University IT, 2024), have rapidly become integral tools for generating marketing and persuasive communication. They are reforming customer experience within the context of marketing by providing more adaptive, personalized, and contextually aware interactions with consumers (Zhou et al., 2026). As a result, organizations now rely on automated text generation from LLMs for email campaigns, promotional announcements, advertising copy, product descriptions, and customer engagement messaging.

When LLMs are used to generate marketing content, they often default to predictable linguistic patterns and emotionless language. As a result, the content that they produce can have the undesirable effect of alienating customers. Thus, it is important to understand the linguistic and emotional characteristics of LLM-generated marketing texts to ensure that they are delivering content to consumers that is authentic, targeted, and engaging.

Despite the widespread adoption of LLM-generated marketing texts, limited empirical research has examined how these models differ in their linguistic and emotional patterns when completing identical marketing tasks; this study was conducted to provide such a comparison.

This study fills a gap in the literature by conducting a controlled, cross-model comparison of linguistic profiles across three major LLMs (GPT-5.1, Gemini, and Perplexity). Using a standardized set of 30 marketing and persuasive prompts, we evaluate whether LLMs differ in emotional tone, semantic framing, and psycholinguistic style, when given identical instructions. By holding prompts constant and analyzing Empath-derived lexical categories across model outputs, we isolate stylistic differences that arise from each model's training data, architecture, or generation tendencies.

Literature Review

Much of the existing literature evaluates LLMs based on accuracy, benchmark performance, reasoning, and fairness (Bubeck et al., 2023). However, comparatively fewer studies have investigated the psycholinguistic profiles of LLM-generated content, particularly in domains where emotional tone, aspirational messaging, cognitive framing, and social orientation directly influence business outcomes. Marketing communication relies heavily on such features, and small linguistic differences may shape consumer perceptions, engagement rates, and persuasive effectiveness.

To analyze linguistic features systematically, past research has often turned to Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015). However, LIWC's lexicon, while robust, is limited to manually curated categories. Modern marketing language often contains newer vocabulary, product terminology, and social-media-oriented semantics that LIWC may not fully capture in addition, LIWC is limited to only 40 topical and emotional categories. To address these limitations, this study uses Empath (Fast et al., 2016), an open-source, neural-embedding–based lexical analysis tool capable of detecting over 190 semantic categories, including optimism, achievement, business, social media, technology, and other domains highly relevant to AI-generated marketing text.

The emergence and widespread use of LLMs is rather new, with most appearing to the general public over the past four years. Thus, the studies conducted on LLMs are rather limited. Specifically, within the context of marketing, several authors have studied the employment of AI. These studies focus on the creation of personalized marketing content, the development of marketing strategies, analysis used in marketing research, and the use of LLMs in marketing management.

Recognizing today's consumer's desire for highly personalized marketing content, Büyüksomer and Tekeoğlu (2024) investigate the use of generative artificial intelligence (GenAI) in various marketing strategies. Their study details the creation of personalized audio, video, and text marketing content. They also investigate the use of chatbots and virtual assistants. Dewivedi et al. (2023) also consider the employment of GenAI in the context of marketing. They find that GenAI has contributed to marketing by building marketing campaigns, the creation of chatbots, the creation of art, and providing analytics on brand comparison, customer experience, marketing research, etc. They weigh the benefits of its use with the risks resulting from the possible use of offensive or copyrighted content, loss of privacy, fraudulent transactions, and spreading of false information.

Huang and Rust (2021) study mechanical AI, thinking AI, and feeling AI. They investigate the ways in which these various forms of AI are able to be employed in marketing research, strategy, and action. They conclude that marketing primarily requires thinking AI and feeling AI and while thinking AI has made rapid progress, feeling AI continues to develop. Qian et al. (2025) investigate the use of LLMs across five distinct areas of marketing including idea and content generation and the substitution for human participants in research. Zhou and Song (2026) study the effect that LLMs are having on the design, delivery, and governance of customer experience. They examine six areas where LLMs are showing considerable impact and conclude that LLMs are playing an important role in defining brand experience and customer relationships.

Brand et al. (2026) investigate the use of LLMs for market research in understanding consumer preferences. They suggest that LLMs should be used to supplement traditional human studies, rather than to replace them. Stromberg et al. (2025) study the use of LLMs to simulate consumer responses to brand tracking surveys by comparing LLM survey results to that of human responses captured by YouGov's BrandIndex. They find mixed results regarding the accuracy of the LLMs responses compared to real consumers. Estevez et al (2025) also studied the use of LLMs by investigating the accuracy of using GenAI in traditional market research. They compared the GenAI results produced from four different LLMs with those of human responses and concluded that although the LLMs could serve as reliable proxies, they cannot fully replicate the results of traditional human marketing research. Recognizing that a gap exists between LLM-generated responses and human responses when performing conjoint analysis, Wang et al. (2025) propose a statistical data augmentation method that combines LLM-generated responses with human responses.

Aghaei, et al. (2025) study the employment of LLMs in marketing management. Their study focuses on the use of LLMs in marketing personalization, real-time-interactive customer insights, and automatic content generation. They aim to provide marketers with guidance on integrating LLMs into their marketing strategy by using industry standard best practices.

Research Questions

The goal of this study is to investigate the linguistic and emotional patterns produced by the GPT-5.1, Gemini Flash 2.5, and Perplexity Sonar LLMs when generating marketing content. Specifically, this study sought to answer the following research questions.

RQ1. Do different LLMs produce systematically different semantic and emotional patterns when generating persuasive marketing content, as measured through Empath?

RQ2. How do LLMs vary in their use of marketing-relevant psycholinguistic categories, such as achievement, optimism, business, and technology, when given identical prompts?

RQ3. Where model-level differences do appear, are they consistent across prompt types (emails, product descriptions, social media posts, announcements), or do they depend on the kind of marketing task?

RQ4. Do some LLMs exhibit stronger tendencies toward persuasive positivity, aspirational framing, or achievement-oriented messaging than others?

RQ5. Do the LLMs differ in the stability of their linguistic style across prompts; that is, within-model variability — independent of any differences in their average linguistic profiles?

Methodology

Research Design

This study employed a controlled, cross-model comparative design to examine linguistic and emotional differences in marketing content generated by LLMs. Identical prompts were submitted to three commercially available LLMs (GPT-5.1, Gemini Flash 2.5, and Perplexity Sonar) under deterministic generation conditions, and the resulting texts were analyzed using Empath, an open-source psycholinguistic tool. By holding prompts and generation parameters constant, the design isolates stylistic and emotional differences attributable to model architecture and training data rather than task variation.

Prompt Corpus

A corpus of 30 standardized marketing and persuasive prompts was constructed to reflect common real-world business communication tasks. Prompt categories included promotional emails, product descriptions, advertising copy, social media posts, calls-to-action, and service announcements. Each prompt was written to be comparable in length, scope, and communicative intent, while varying surface topic to reduce prompt-specific bias. Since the corpus was developed by the authors, this study should be interpreted as exploratory research using a preliminary, researcher-developed prompt corpus intended to support an initial cross-model comparison under standardized conditions, not as a confirmatory evaluation based on a fully validated prompt instrument. This is a recognized approach per Arain, Campbell, Cooper & Lancaster (2010), The full prompt list is provided in Appendix A.

Language Models Evaluated

Three commercially available LLMs that are currently being used to generate marketing content were evaluated: GPT-5.1, Gemini Flash 2.5, and Perplexity Sonar. OpenAI's GPT-5.1 LLM boasts an improvement on chat conversation when compared to previous GPT versions (OpenAI, 2025). Gemini is the family of LLMs developed by Google DeepMind; it promises strong generalist capabilities and cutting-edge understanding and reasoning skills (Google DeepMind, 2025). Perplexity AI's LLM, Perplexity, employs retrieval-augmented generation (RAG), combining the pre-trained knowledge of traditional LLMs with up-to-date external data that is fetched from outside sources in real time (Perplexity AI, 2024).

Perplexity was deliberately included despite operating as a RAG system rather than a purely generative model. This decision reflects the reality that end-users increasingly treat RAG-based and traditional LLM platforms interchangeably when generating marketing content, and a practical comparison should therefore include the tools practitioners actually use. Perplexity's underlying model is not publicly specified in a way that enables exact architectural replication in this study. Its outputs may incorporate web-sourced content, which could introduce web-grounded phrasing, citation behavior, or stylistic patterns that are not present in purely generative systems. This architectural difference is acknowledged as a methodological consideration and is addressed in the interpretation of results and in the Limitations section.

Each model generated one response per prompt. To maximize reproducibility and reduce stochastic variation, all generations were produced using near-deterministic settings (temperature = 0), although minor variability may persist, especially in retrieval-augmented systems. A maximum output limit of approximately 250 tokens was specified. However, all three models consistently produced substantially shorter outputs, indicating that response length was governed by prompt completion rather than by the token limit. Despite the standardized upper bound, output lengths varied significantly across models: GPT-5.1 averaged 32.1 words per output (SD = 5.5), Gemini averaged 34.8 words (SD = 6.1), and Perplexity averaged 49.6 words (SD = 8.6). A one-way ANOVA confirmed this difference was highly significant, $F(2, 87) = 57.0$, $p < 0.0001$. This word count disparity is addressed in the analysis.

Data Collection and Corpus Construction
For each prompt–model pair, the generated output was recorded and stored in a structured dataset containing prompt id, prompt text, model, and output text. This process produced a corpus of 90 documents (30 prompts × 3 models). All outputs were generated within a short time window to minimize version drift in deployed models.

Psycholinguistic Feature Extraction Using Empath
Each generated text was processed through Empath (Fast et al., 2016), an open-source lexical analysis library built on neural word embeddings that is able to analyze text across categories and generate and validate new lexical categories on demand. Empath provides normalized scores for over 190 validated semantic and affective categories, including emotional processes (e.g., positive emotion, joy, fear, and trust), motivational and social themes (e.g., achievement, power, optimism, and social media), and economic and business language (e.g., money, business, technology, and economics). Scores represent the proportion of words in each text matching each category's lexicon, thereby normalizing for text length.

Variable Selection
From the full Empath output, 15 focal categories were selected based on their theoretical relevance to marketing communication: positive emotion, negative emotion, joy, trust, fear, optimism, achievement, power, social media, art, money, business, work, technology, and economics. These categories reflect a composite of emotional tone, persuasive framing, business orientation, and technological positioning. An exploratory scan of all 194 Empath categories was also conducted to identify any significant differences beyond the set.

Statistical Analysis
Prior to inferential testing, a floor-effect check was conducted to identify categories with insufficient variance. Categories in which two or fewer of the 90 texts produced non-zero scores were excluded from inferential analyses and reported descriptively only. This threshold was chosen because inferential testing is statistically unstable and uninformative under such extreme sparsity, where nearly all observations are zero.

For testable categories, the following analyses were performed: (a) one-way analysis of variance (ANOVA) for each category, with model as the independent variable; (b) Bonferroni and Benjamini–Hochberg (BH) corrections for multiple comparisons, applied across all testable ANOVAs simultaneously; (c) Kruskal–Wallis tests as nonparametric robustness checks; (d) Tukey HSD and pairwise Mann–Whitney U tests with Cohen's d effect sizes for categories reaching the $p < .10$ threshold; (e) multivariate analysis of variance (MANOVA) to test for overall multivariate differences; (f) analysis of covariance (ANCOVA) controlling for word count; and (g) cross-prompt stability analysis using coefficients of variation. All analyses were conducted in Python using Empath, SciPy, statsmodels, and scikit-learn. All results were thoroughly reviewed by the authors for accuracy, and all code is available by request. It is not included due to space

limitations. The cross-prompt stability analysis, computed as the coefficient of variation, addresses RQ5 by quantifying within-model variability independently of between-model differences in mean scores.

Results

This section presents the full set of empirical findings comparing linguistic and emotional characteristics across the three LLMs based on the 15 focal Empath categories previously discussed and the full 194-category Empath output.

Floor-Effect Check

Of the 15 focal categories, three exhibited floor effects that rendered inferential testing inappropriate. Negative emotion produced non-zero scores in only one of 90 texts, while joy and fear produced non-zero scores in zero of 90 texts. These categories were excluded from the ANOVA and reported descriptively only. The remaining 12 categories were deemed testable and carried forward for statistical analysis which is shown in Table 1. Categories labeled as sparse (those having three through 19 non-zero texts) should be interpreted with caution, as low base rates reduce statistical power and stability. Categories with at least 20 non-zero texts were considered adequate. Status is a data-sufficiency rating noting which categories carried enough signal to trust, and justifies why three were excluded from the MANOVA/ANOVA and why some "testable" results still warrant caution

Table 1. Floor-Effect Check for 15 Focal Empath Categories

Category	Non-Zero Texts (of 90)	Status
Negative emotion	1	Untestable
Joy	0	Untestable
Fear	0	Untestable
Power	8	Sparse
Economics	9	Sparse
Positive emotion	14	Sparse
Achievement	16	Sparse
Optimism	18	Sparse
Art	21	Adequate
Trust	22	Adequate
Money	31	Adequate
Social media	35	Adequate
Work	53	Adequate
Technology	53	Adequate
Business	58	Adequate

Descriptive Statistics

Table 2 presents the mean Empath scores for each of the 15 focal categories across the three LLMs. The values presented within the table represent the proportion of word matches to total words, which have been normalized by Empath. As shown in the table, business, work, and technology were the most active categories across all models, consistent with the marketing domain. Emotional categories (joy, fear, and

negative emotion) were near zero for all models, indicating uniformly positive and emotionally restrained marketing text.

Table 2. Mean Empath Scores for 15 Focal Categories by Model

Category	GPT-5.1	Gemini	Perplexity
Positive emotion	0.0074	0.0039	0.0020
Negative emotion	0.0013	0.0000	0.0000
Joy	0.0000	0.0000	0.0000
Trust	0.0057	0.0019	0.0021
Fear	0.0000	0.0000	0.0000
Optimism	0.0085	0.0054	0.0034
Achievement	0.0032	0.0088	0.0006
Power	0.0009	0.0050	0.0007
Social media	0.0060	0.0060	0.0077
Art	0.0049	0.0084	0.0059
Money	0.0034	0.0047	0.0026
Business	0.0191	0.0164	0.0228
Work	0.0115	0.0171	0.0186
Technology	0.0164	0.0225	0.0137
Economics	0.0070	0.0078	0.0082

MANOVA – Overall Multivariate Differences

A MANOVA was conducted using the 12 testable categories to determine whether the three LLMs differed in their combined semantic profiles. The results indicated no statistically significant overall differences: Wilks' Lambda = 0.630, $F(28, 148) = 1.375$, $p = 0.116$; Pillai's Trace = 0.420, $F(28, 150) = 1.422$, $p = 0.093$. This suggests that, when considered simultaneously, the three LLMs produce broadly similar marketing-related language profiles.

Univariate ANOVA with Multiple Comparisons Corrections

One-way ANOVAs were conducted for each of the 12 testable Empath categories. To control for inflated Type I error, both Bonferroni and Benjamini–Hochberg (BH) corrections were applied across all 12 tests. The results are reported in Table 3. After correction, no categories reached statistical significance at $\alpha = 0.05$. One category, achievement, which is highlighted in the table, approached significance at the uncorrected level ($F(2, 87) = 3.69$, $p = 0.029$, $\eta^2 = 0.078$) but did not survive Bonferroni correction ($p_{corrected} = 0.348$) or Benjamini–Hochberg correction ($p_{BH} = 0.348$). The remaining adjusted BH values ranged from 0.654 to 0.975.

Table 3. One-Way ANOVA Results with Multiple Comparisons Corrections (12 Testable Categories)

Category	F	p (unc.)	p (Bonf.)	p (BH)	η^2	n > 0
Achievement	3.689	0.029*	0.348	0.348	0.078	16
Power	1.851	0.163	1.000	0.654	0.041	8
Positive emotion	1.566	0.215	1.000	0.654	0.035	14
Business	1.551	0.218	1.000	0.654	0.034	58
Trust	0.652	0.524	1.000	0.975	0.015	22
Art	0.657	0.521	1.000	0.975	0.015	21
Economics	0.538	0.586	1.000	0.975	0.012	9
Work	0.339	0.713	1.000	0.975	0.008	53
Social media	0.168	0.846	1.000	0.975	0.004	35
Technology	0.131	0.877	1.000	0.975	0.003	53
Optimism	0.104	0.901	1.000	0.975	0.002	18
Money	0.025	0.975	1.000	0.975	0.001	31

Nonparametric Robustness Check

Kruskal–Wallis tests were conducted for all 12 testable categories. Consistent with the parametric results, no category reached statistical significance at $\alpha = 0.05$, confirming that the null finding is robust to distributional assumptions.

Exploratory Post-Hoc Analysis

Although no category reached corrected significance, exploratory post-hoc analyses (Tukey HSD and pairwise Mann–Whitney U with Cohen's d) were conducted for the top categories. Gemini produced descriptively higher achievement scores (M = 0.0088) than GPT-5.1 (M = 0.0032) and Perplexity (M = 0.0006). However, pairwise comparisons did not reach significance after accounting for family-wise error, and effect sizes were in the small-to-medium range.

ANCOVA - Controlling for Word Count

As shown in Table 4, given the significant differences in output length ($F(2, 87) = 57.0$, $p < 0.0001$), ANCOVA was conducted for each testable category with word count as a covariate. For the majority of categories, controlling for word count weakened the already non-significant model effects, suggesting that some descriptive differences may be partially attributable to output length rather than genuine stylistic variation.

Table 4. ANCOVA Summary

Analysis	Reported Result	Interpretation
ANCOVA with word count covariate	For the majority of categories, controlling for word count weakened the already non-significant model effects.	Observed descriptive differences were partially attributable to output length rather than clear stylistic differences.
Word count effect	Output length differed significantly across models, $F(2, 87) = 57.0$, $p < 0.0001$.	Perplexity produced longer outputs, making length an important confound to consider in interpretation.

Full 194-Category Exploratory Scan

An exploratory ANOVA scan across all testable Empath categories from the full 194-category output was conducted with Benjamini–Hochberg correction. No categories reached statistical significance after correction, confirming that the convergence extends to the full breadth of Empath's lexicon.

Cross-Prompt Stability

As shown in Table 5, the coefficient of variation (CV) was computed for each model across all 12 testable categories. While all three of the models exhibited an extremely high cross-prompt variability, Gemini exhibited the most consistent style (mean CV = 2.22), followed by GPT-5.1 (mean CV = 2.50), with Perplexity showing the most variability (mean CV = 2.67). Perplexity's higher variability is consistent with its retrieval-augmented architecture, which may introduce content-dependent variation from web sources. These results indicate that the models differ reliably in stability even though they do not differ in their average profiles, answering RQ5 in the affirmative: variability, unlike central tendency, distinguishes the models.

Table 5. Cross-Prompt Stability: Coefficient of Variation by Model and Category

Category	GPT-5.1	Gemini	Perplexity
Achievement	3.04	1.98	5.39
Art	2.28	1.98	2.43
Business	1.62	1.48	0.97
Economics	2.55	2.69	1.88
Money	4.04	3.48	3.24
Optimism	1.66	2.04	2.66
Positive emotion	2.12	2.55	3.01
Power	5.39	2.27	5.39
Social media	2.25	2.00	1.46
Technology	1.16	1.18	1.21
Trust	2.59	3.74	3.02
Work	1.33	1.30	1.37
Mean CV	**2.50**	**2.22**	**2.67**

Summary of Findings

No statistically significant multivariate or univariate differences were found after correction for multiple comparisons, and the full 194-category exploratory scan confirmed this result. Three emotional categories exhibited floor effects. Output length varied significantly, with Perplexity producing approximately 50% more words. Cross-prompt stability analysis indicated Gemini is the most stylistically consistent model. Relative to the research questions, these results can be summarized as follows:

- **RQ1:** The three LLMs did not produce systematically different semantic or emotional patterns; after correction, no category differed significantly across models.
- **RQ2:** Model-level variation in marketing-relevant categories (achievement, optimism, business, technology, and others) was minimal, with all effect sizes small ($\eta^2 < .08$).
- **RQ3:** Because no reliable model-level differences emerged, none proved task-specific; the convergence held uniformly across emails, product descriptions, social media posts, and announcements.

- **RQ4:** No model showed a statistically reliable tendency toward persuasive positivity or achievement-oriented messaging; the descriptive achievement trend (Gemini highest) did not survive correction.
- **RQ5:** The models did differ in the stability of their style across prompts, independent of their average profiles, with Gemini the most consistent (mean CV = 2.22) and Perplexity the least (mean CV = 2.67).

Discussion

This study asked whether different LLMs produce different linguistic and emotional patterns when generating marketing content. After correcting for multiple comparisons, no significant differences emerged across the 15 focal Empath categories or the full 194-category lexicon.

We suggest this convergence as evidence of a specific cause, not simply an absence of differences. If architecture and training data were what mainly shaped marketing style, three models built by three companies on different data should each leave a distinct stylistic mark. Our results show they do not. What GPT-5.1, Gemini, and Perplexity have in common is not how they were pretrained but how they were aligned. Instruction tuning and reinforcement learning from human feedback (RLHF) seem to push all of them toward the same goal: output that reads as professional, helpful, and persuasive. That shared goal appears to pull different models toward a common style.

This explanation also predicts where the similarity should break down. Alignment should dominate only when the task is tightly constrained This would include standardized prompts, little ambiguity, and deterministic settings, as in this study. If more open-ended prompts that call for a distinctive brand voice were provided, differences should appear. Our finding suggests that when provided with narrow constraints, results from the three tools are similar. Note that this is still an exploratory study and generalizations to a broader corpus are not warranted. Rather, our study serves as a starting point for a deeper understanding of the variances between various models. Two smaller findings support this. Achievement was the only category to approach significance before correction ($p = .029$, with Gemini highest and Perplexity lowest). Perplexity's output length is approximately 50% greater despite identical token limits; this finding held consistent between-model differences. However, the gap comes from architecture, not style: Perplexity's retrieval system pulls in web content that lengthens and varies its output. Therefore, where the models diverge, they diverge on a structural feature.

The emotional floor effects make a similar case. All three models suppress joy, fear, and negative emotion, producing uniformly upbeat, emotionally flat text. Three separately trained systems avoiding the same emotional extremes are hard to equate to chance and can be explained through safety-oriented alignment. The models agree not only on what to say but on what to leave out. The practical takeaway is that if you want emotional variety it must be provided through prompting.

Finally, similar averages do not mean similar consistency. Gemini kept the most stable style across prompt types (mean CV = 2.22) and Perplexity the least (mean CV = 2.67), an ordering that, again, tracks architecture rather than style. Consistency is where the models remain genuinely distinguishable even when their average profiles look alike. Even though the three models have similar averages (their typical style is alike), they don't have similar consistency (some stick close to that typical style, others are more varied). Gemini stays close to its average from prompt to prompt, while Perplexity is more varied. So, the average makes them look interchangeable, but the consistency measure still differentiates them. This matters for anyone using these tools, since the model chosen has little effect on the overall style, the real control comes

from other choices: how you write the prompt, what brand-voice rules you set, and how you edit the text afterward. The model still matters, albeit for narrower reasons such as output length, how it pulls in information, and how consistent it is, rather than its tone or persuasive power. When steady, consistent writing is the goal, a model like Gemini has an edge. When the goal is a unique or emotionally rich voice, no model provides that on its own; therefore, a strong prompt must be used.

Limitations

The dataset that was used for this study consisted of 30 prompts per model (90 texts total), which may lack the power to detect subtle differences. The small effect sizes observed (all $\eta^2 < .08$) suggest that larger samples would be needed to detect reliable model-level differences.

Empath is a lexicon-based tool relying on surface-level word matching. It does not capture context, irony, or rhetorical strategies. Transformer-based embedding methods (e.g., Sentence-BERT or BERTopic) could provide richer contextual semantics in future work.

Only three LLMs were examined. The rapidly evolving ecosystem includes many additional models that may exhibit different characteristics. In addition, due to the speed of AI evolution, there are already newer versions of the three LLMs on the market. Each new iteration of an LLM boasts the deployment of better understanding, reasoning, and content generation.

Although Perplexity was deliberately included because end-users commonly treat RAG-based and traditional LLM platforms interchangeably, its retrieval-augmented architecture introduces a structural asymmetry. Its longer outputs and higher variability may stem from web-sourced content rather than from differences in generative capability, and readers should weigh cross-model comparisons involving Perplexity accordingly.

The significant word count differences across models represent a confound that could not be fully eliminated despite ANCOVA sensitivity checks. More specifically, Perplexity produced approximately 50% more text than the other models despite identical token limits. Given the relatively short length of generated texts (approximately 30–50 words), Empath scores may be sensitive to individual word occurrences, which can amplify variance and limit the stability of lexical estimates. ANCOVA showed that controlling for word count weakened model effects, suggesting some descriptive differences may reflect length-related artifacts. Future studies should consider tighter length constraints or length-matched analyses.

As noted previously, since the corpus was developed by the authors, this study should be interpreted as exploratory research using a preliminary, researcher-developed prompt corpus intended to support an initial cross-model comparison under standardized conditions, not as a confirmatory evaluation based on a fully validated prompt instrument. Finally, the generic marketing prompts may not reflect industry applications involving tailored brand guidelines, jargon, or regulatory constraints. This study compares model outputs, not the internal capabilities of the models themselves. Differences in output style may reflect training data composition, alignment tuning, or decoding strategies rather than fundamental architectural advantages.

Conclusion

This study provides one of the first rigorous, correction-adjusted empirical comparisons of linguistic and emotional characteristics across multiple LLMs generating identical marketing texts. Using Empath to quantify 194 semantic categories and applying ANOVA with Bonferroni and Benjamini–Hochberg corrections, MANOVA, Kruskal–Wallis tests, and ANCOVA, we find that GPT-5.1, Gemini, and Perplexity produce broadly convergent marketing language profiles under deterministic prompting conditions, with no statistically significant differences after proper correction.

This convergence is the study's central contribution. It suggests that under these conditions, contemporary LLMs, shaped by similar training paradigms and alignment objectives, arrive at remarkably similar linguistic output when given identical marketing tasks. Organizations evaluating LLMs for marketing content should not expect dramatically different linguistic characteristics or brand voices based solely on model choice. Prompt quality, brand voice guidelines, and post-generation editing are likely to have greater impact than model selection when creating desired marketing communications. The cross-prompt stability findings suggest Gemini may be preferable for applications requiring a consistent tone across communication outlets.

The absence of joy, fear, and negative emotion reflects a consistent characteristic of LLM-generated marketing text: it is uniformly positive and emotionally restrained. This is likely a product of safety-oriented training and RLHF. Organizations seeking emotionally differentiated content may need explicit prompt engineering to elicit non-default emotional tones.

Future work should expand the model set, increase corpus size, incorporate embedding-based methods, test non-deterministic settings, and explore whether more targeted prompts can elicit model-level differences that generic prompts do not. As LLMs continue to shape digital communication, understanding their linguistic convergence remains essential for ethical and effective AI-assisted marketing.

Acknowledgments

AI Use Declaration: During the preparation of this work the author(s) used Anthropic Claude in order to assist in statistics, graphics development, and content review. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

Aghaei, R., Kiaei, A. A., Boush, M., Vahidi, J., Zavvar, M., Barzegar, Z., & Rofoosheh, M. (2025). Harnessing the Potential of Large Language Models in Modern Marketing Management: Applications, Future Directions, and Strategic Recommendations. *ArXiv, abs/2501.10685.*

Arain, M., Campbell, M. J., Cooper, C. L., & Lancaster, G. A. (2010). What is a pilot or feasibility study? A review of current practice and editorial policy. BMC medical research methodology, 10(1), 67.

Brand, J., Israeli, A., & Ngwe, D. (2026). *Using LLMs for market research*. (Working Paper No. 23-062) . Harvard Business School. https://www.hbs.edu/ris/Publication%20Files/23-062_1f58623a-ee21-44b9-a262-276047bc5543.pdf

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y., Li, Y., Lundberg, S., Nori, H., Palangi, H., Riberiro, M., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Büyüksomer, A.E., & Tekeoğlu, A. N. T. (2024). Enhancing marketing communications with generative AI: A systematic literature review. Journal of Industrial Policy and Technology Management, 7(2), 109–122.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

Estevez, M., Ballestar, M. T., & Sainz, J. (2025). Market research and knowledge using Generative AI: the power of Large Language Models. Journal of Innovation & Knowledge 10(5).

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: understanding topic signals in large-scale text. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 4647–4657.

Google DeepMind. (2025). Gemini: a family of highly capable multimodal models. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf#page=71

Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. Journal of the Academy of Marketing Science, 49, 30–50.

OpenAI. (2025). GPT-5.1: a smarter, more conversational ChatGPT. https://openai.com/index/gpt-5-1/

Pennebaker, J. W., Boyd, R. L., Jordan, K. N., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin. DOI: 10.15781/T29G6Z

Perplexity AI. (2024). An introduction to RAG models. https://www.perplexity.ai/page/an-introduction-to-rag-models-jBULt6_mSB2yAV8b17WLDA

Qian, Y., Chen, S., Wu, C., Yuan, K., Du, Y., Jiang, Y., & Liu, Y. (2025). Large Language Models for Marketing Research: A Survey and New Perspectives. Proceedings of the 58th Hawaii International Conference on System Sciences, 1249–1259.

Standford University IT. (2024, December, 13). *AI Demystified: Introduction to large language models* [Online training]. https://uit.stanford.edu/service/techtraining/ai-demystified/llm

Stromberg, M., Moe, W., Reutterer, T., & Schweidel, D. (2025). Blind Spots in Broad Strokes: Caveats for the Use of LLMs in Marketing Research. Available at SSRN: https://ssrn.com/abstract=5145114 or http://dx.doi.org/10.2139/ssrn.5145114

Wang, M., Zhang, D., & Zhang, H. (2026). Large Language Models for Market Research: A Data-Augmentation Approach. Marketing Science 0(0). https://doi.org/10.1287/mksc.2025.0009

Zhou, W. & Song, Y. (2026). How LLMs are redefining customer experience in marketing. In: Ladhari, R. (eds) Encyclopedia of Artificial Intelligence in Marketing. Springer, Cham. https://doi.org/10.1007/978-3-031-75316-9_71-1

Appendix

Standardized Marketing Prompt Set

Each prompt was designed to elicit short-form marketing communication under comparable business conditions.

1. Write a promotional email for a new productivity app.
2. Create a product description for a smartwatch.
3. Write a social media post announcing a seasonal sale.
4. Draft a marketing email for a fitness subscription service.
5. Create a product description for wireless earbuds.
6. Write a promotional post for a new coffee shop opening.
7. Draft a call-to-action for an online course platform.
8. Create an advertisement for a travel booking website.
9. Write a product description for an electric vehicle.
10. Draft a promotional email for a clothing brand.
11. Write a social media post promoting a mobile banking app.
12. Create a product description for a home security system.
13. Draft an announcement for a software product launch.
14. Write a promotional message for a streaming service.
15. Create a product description for a gaming console.
16. Write a marketing email for a meal delivery service.
17. Draft a social media campaign post for a skincare product.
18. Create a product description for a laptop computer.
19. Write a promotional email for a hotel chain.
20. Draft an advertisement for a ride-sharing service.
21. Create a product description for a smart home assistant.
22. Write a social media post for a limited-time discount.
23. Draft a promotional email for an online marketplace.
24. Create an advertisement for a financial planning service.
25. Write a product description for a fitness tracker.
26. Draft a marketing email for a language learning app.
27. Create a promotional post for a music streaming service.
28. Write a product description for a digital camera.
29. Draft an announcement for a new restaurant opening.
30. Write a call-to-action for a subscription-based newsletter.