Marketing Science Institute Working Paper Series 2025

Report No. 25-127

Assessing Artificial Marketing Intelligence

Raymond R. Burke, Maximilian Matthe and Alex Leykin

Assessing Artificial Marketing Intelligence

Raymond R. Burke*
E.W. Kelley Professor of Business Administration
Kelley School of Business
Indiana University
Hodge Hall 2100, 1309 E. 10th St.,
Bloomington, IN 47405, United States
rayburke@iu.edu

Maximilian Matthe*
Assistant Professor of Marketing
Kelley School of Business
Indiana University
Hodge Hall 2100, 1309 E. 10th St.,
Bloomington, IN 47405, United States
mpmatthe@iu.edu

Alex Leykin
DataBelay LLC
databelay@gmail.com

Declarations of interest: none

* The first two authors contributed equally and are represented in alphabetical order.

We thank conference participants at the 2024 MarkTech Conference and the 2025 AIM Conference for their comments that helped improve the paper. We also thank Madhura Ashtekar and Nakul Havaldar for their help in implementing this project.

Assessing Artificial Marketing Intelligence

Abstract

Foundational marketing knowledge, from theories of consumer behavior to segmentation frameworks and pricing models, supports effective marketing decisions, yet managers often struggle to access and apply it in practice. This paper evaluates whether large language models (LLMs) can function as scalable marketing knowledge systems that make such knowledge broadly accessible. Specifically, we assess what LLMs "know" about marketing and how effectively they can reason with and apply that knowledge. We construct a comprehensive dataset of over 30,000 questions drawn from instructor materials of 25 widely used marketing textbooks and assess LLM performance across three dimensions: domain knowledge, reasoning capabilities, and user interaction. Across models and providers, current LLMs achieve high accuracy (84%-94%), demonstrate near-complete coverage of marketing topics, and show strong performance on recall and conceptual understanding. Performance declines on higher-order reasoning tasks, though newer reasoning models close these gaps substantially. Experimental manipulations of these questions suggest that performance reflects conceptual understanding rather than simple memorization. In a human benchmarking study, LLMs substantially outperform respondents across all training levels, while showing complementary strengths relative to more advanced human reasoning. These findings indicate that LLMs have crossed an important capability threshold: they now provide reliable, on-demand access to much of the discipline's foundational knowledge, with important implications for how marketing knowledge is disseminated, taught, and applied.

Keywords: artificial intelligence, large language models, knowledge systems, marketing knowledge

1 Introduction

Informed marketing decisions depend on access to and effective use of marketing knowledge. One important type is *marketplace knowledge*: tacit, experience-based insights that are often grounded in market research and focused on customers, competitors, and firm-specific contexts. The other is *foundational knowledge* about marketing itself: the systematically developed body of validated beliefs about key marketing constructs, principles, and frameworks (Rossiter, 2001), which are codified in textbooks and taught in business schools. Marketplace knowledge concerns the who, what, when, and where of marketing (the "Five Cs"), while foundational marketing knowledge addresses the how and why of marketing (e.g., the underlying mechanisms of the communication process and customer decision-making; strategies for segmentation, targeting, and positioning; approaches for brand equity management). Used together, these two types of knowledge enable managers to make more informed and grounded decisions.

While the challenges and benefits of acquiring and applying marketplace knowledge have received extensive attention in the strategy literature (e.g., Glazer, 1991; Day, 1994; Moorman & Miner, 1997; Winter, 2009; Germann et al., 2013), foundational marketing knowledge has received far less scrutiny. Yet it plays a fundamentally important role in business judgment, as it allows managers to make decisions based on established principles rather than experience or intuition alone. In many business contexts, foundational marketing knowledge can have a profound influence on managerial decision-making (Roberts et al., 2014).

Often, however, foundational marketing knowledge remains a scarce and inaccessible resource. Today's canon of marketing knowledge spans a vast array of theories, models, concepts, and empirical findings (e.g., Eisend, 2015; Wierenga, 2021). For firms, acquiring this knowledge requires costly investments, such as recruiting formally trained talent, sponsoring continuing education, commissioning consultants, or conducting original research, which can be particularly challenging for smaller or resource-constrained firms. Even then, managers may struggle to retrieve the relevant parts of this knowledge at the point of decision-making and instead rely on personal

experience, analogy, or intuition (e.g., Wierenga & Van Bruggen, 1997; Wierenga, 2002). A recent survey by Ipsos (2026) revealed that many marketing professionals fail basic assessments of marketing fundamentals. This challenge contributes to a broader gap between academic research and practice. It is frequently noted that the ivory tower of academia can be too distant from the practicalities of the real world and thus fails to *generate* relevant research (e.g., Lilien et al., 2013; Kohli & Haenlein, 2021b,a; Stremersch, 2021; Van Heerde et al., 2021; Wierenga, 2021; Schauerte et al., 2023). Yet another substantial challenge is finding effective channels for *disseminating* foundational knowledge to marketing professionals (e.g., Lehmann, 2014; Kumar, 2017; AMA, 2025). Despite various efforts to do so, foundational marketing knowledge remains hard to access and apply in practice, which limits the practical impact of academic research.

This paper argues—and empirically demonstrates—that generative AI, particularly large language models (LLMs), can offer a new path forward by making the established body of foundational marketing knowledge widely accessible. We conceptualize LLMs as a new form of knowledge system (e.g., Rangaswamy et al., 1989; Burke et al., 1990; Wierenga, 1990; Burke, 1991) capable of storing, retrieving, and applying domain-specific knowledge. Trained on vast corpora that include academic texts and scholarly publications, LLMs encode a representation of the foundational knowledge about our discipline and offer practitioners a simple interface to access it during decision-making. Thus, as AI systems become more deeply embedded in marketing functions (Deveau et al., 2023; Korst et al., 2024), they have the potential to transform foundational marketing knowledge from a scarce and unevenly distributed resource into a more widely accessible asset—with far-reaching implications for research, education, and practice.

Realizing this potential, however, hinges upon the question: *How much do LLMs truly know about marketing?* While substantial research has examined what LLMs *can do* in marketing, that is, their functional ability to perform various tasks, no prior work has systematically evaluated what LLMs *know* about marketing: the extent to which they correctly encode marketing concepts, frameworks, principles, and methods (Rossiter, 2001), and how effectively they can apply this knowledge in reasoning and decision-making. Although LLMs often produce plausible and confi-

dent responses, existing evidence remains limited to individual cases (e.g., Jürgensmeier & Skiera, 2024). Particularly in a field like marketing, which intersects a variety of disciplines (e.g., statistical analyses, strategic decision-making, consumer psychology) and demands varied cognitive skills, isolated pieces of evidence do not offer sufficient validation to establish whether LLMs possess a reliable representation of the entire marketing knowledge domain.

Moreover, known limitations of LLMs suggest caution. LLMs are trained on vast online content, including superficial or inconsistent marketing advice from blogs, self-proclaimed experts, or promotional material—possibly disconnected from established academic perspectives. Their capabilities also exhibit "jagged frontiers," with strong performance in one area coexisting with unexpected failures in adjacent ones (e.g., Dell'Acqua et al., 2026; Karpathy, 2024; Saxena et al., 2025). Rather than applying structured reasoning, LLMs may operate as surface learners, identifying patterns without understanding the deeper conceptual logic (Zellers et al., 2019). In addition, prompt sensitivity may further undermine their outputs' reliability (e.g., Mohammadi, 2024; Brucks & Toubia, 2025). Therefore, we argue that a rigorous, systematic evaluation of LLMs' marketing knowledge, reasoning abilities, and possible limitations in human-AI interactions is necessary.

This paper offers the first such evaluation. We empirically assess LLMs' marketing knowledge by testing their ability to answer questions drawn from a core academic corpus. Specifically, we compile a dataset of over 30,000 questions derived from the supplementary materials of 25 marketing textbooks widely used across undergraduate and graduate curricula in the US. We then assess LLMs' performance on this corpus to determine the accuracy, sensitivity, and robustness of their marketing knowledge.

Our analysis focuses on three dimensions, building on earlier work on knowledge-based systems (e.g., Burke et al., 1990): (1) whether LLMs possess relevant domain knowledge, (2) whether they can reason effectively with that knowledge, and (3) how their performance varies across different forms of user interactions. To assess LLMs' domain knowledge, we evaluate their performance on questions across and within different marketing topics. To assess their reasoning abilities, we evaluate their performance as a function of cognitive complexity. To assess the impact of user

interactions, we evaluate LLM performance across question-wording (i.e., linguistic features of the question) and alternative prompt designs. To address concerns that such an evaluation may just test LLMs' abilities to recall answers from their training data, we exploit residual within-textbook variation in our regression analyses and conduct experimental manipulations to test whether LLMs can generalize beyond surface-level cues. We further extend our findings to an open-ended question format, and benchmark results against human marketers at different levels of training and experience.

Results reveal strong and improving performance. As of Spring 2026, accuracy on 32,990 multiple-choice and true/false questions ranges from $\sim$84% to $\sim$94% across models from major providers (OpenAI, Anthropic, Google, Meta), up from 72.6% in 2023 (GPT-3.5). Accuracy is consistently high across marketing topics, with no systematic knowledge gaps. LLMs recall and understand facts or concepts with near-perfect accuracy, while performance declines moderately for higher-order reasoning tasks (e.g., application, analysis, or evaluation, Krathwohl, 2002). Newer model generations are steadily narrowing these gaps, and reasoning models close them further still. Accuracy is largely stable across variations in question wording and prompt design.

Notably, experimental manipulations suggest that LLMs are not merely reproducing memorized content. Changing question phrasing or answer order reduces accuracy only slightly (-2% to -5%), and restricting the analysis to textbooks with no evidence of answer memorization leaves results unchanged. Results also generalize to open-ended discussion questions, where LLMs from the GPT-4o generation achieve average scores of 78–83 out of 100, validated by both automated and independent human evaluation.

In a human benchmarking study with more than 300 participants who provided nearly 5,000 responses, all respondent groups, from undergraduates (55%) to MBA professionals (61%) and PhD students (63%), fall well below LLM accuracy. LLMs provide substantially greater on-demand access to marketing knowledge than humans at any level of training. Suggestively, as humans advance in their marketing education, their performance profile increasingly diverges from that of LLMs: more experienced respondents excel at higher-order reasoning relative to recall, different from most tested LLMs, pointing to a natural complementarity.

Together, these findings indicate that LLMs have crossed a meaningful threshold in their capabilities and can serve as reliable marketing knowledge systems. In effect, they commoditize access to foundational marketing knowledge, with consequences for research, practice, and education.

Our contribution is twofold. Methodologically, we develop the first systematic benchmark for evaluating LLMs' marketing knowledge. We compile a corpus of over 30,000 questions spanning 12 marketing topics across different levels of cognitive complexity, and develop an evaluation framework that assesses accuracy, sensitivity, and robustness. Beyond the findings we report here, this benchmark can serve as a reusable instrument: as new models are released, the discipline can track whether AI capabilities improve, stagnate, or regress across specific knowledge domains and cognitive demands. Equally, as the marketing canon itself evolves, the benchmark can be updated to extend coverage or assess whether LLMs keep pace with new findings and frameworks.

Substantively, we contribute to the ongoing debate about how AI will transform marketing. While there is ample discussion about how these innovations will impact everyday tasks—such as personalizing ads, conducting secondary research, and performing creative activities (e.g., Ma & Sun, 2020; Peres et al., 2023; Grewal et al., 2025)—we offer the first systematic evaluation of what LLMs *know* about marketing, as distinct from what they can *do*. Our results show that general-purpose LLMs already encode a highly accurate representation of foundational marketing knowledge, which positions them as modern-day knowledge systems (Rangaswamy et al., 1989) and opens a new channel through which academic research can reach practice.

Our findings carry implications for marketing scholarship, where they open new pathways for disseminating academic knowledge into practice; for marketing practice, where LLMs can equalize access to foundational knowledge; and for marketing education, where they provide an empirical basis for rethinking marketing curricula. At the same time, they introduce new challenges around developing marketing expertise, as well as transparency, agency, and control.

In the following sections, we review recent work on the application of large language models and earlier work on building knowledge-based expert systems and discuss the challenges of benchmarking the marketing knowledge of LLMs. Next, we detail our research approach and

describe how we constructed a comprehensive database of marketing questions and answers and then evaluated LLMs' performance along the three dimensions of knowledge, reasoning, and human-AI interactions. We then present the results of our regression-based analysis, experimental manipulations, human benchmarking, and validations with open-ended questions. We close with a general discussion of our findings' implications for academic research, education, and marketing practice, and outline avenues for further research.

2 Background

Our research builds on and extends three streams of literature: (1) GenAI applications in marketing, (2) knowledge systems in marketing, and (3) AI benchmarking. A structured overview of related work is provided in the Web Appendix A.

2.1 GenAI Applications in Marketing

A growing body of work examines how LLMs and other GenAI systems can support and empower marketing stakeholders (Hermann & Puntoni, 2025). Recent studies highlight their effectiveness in generating content for search marketing (Reisenbichler et al., 2022, 2026), improving customer service interactions (Brynjolfsson et al., 2025), replacing human participants in market research (Argyle et al., 2023; Brand et al., 2023; Li et al., 2024), eliciting human preferences (Goli & Singh, 2024), simulating human behavior (Gui & Toubia, 2023; Toubia et al., 2025), and integrating GenAI interactions into marketing research designs (Arora et al., 2025; Joerling, 2026). Other work explores GenAI applications in personalized video production (Kapoor & Kumar, 2025), product design (Burnap et al., 2023), brand logo development (Dew et al., 2022), and visual content generation (Hartmann et al., 2024; Heitmann et al., 2025).

Collectively, these studies show that GenAI performs well on a wide range of marketing tasks. However, most work focuses on specific marketing tasks (such as generating ad copy or simulating respondents), which are quite narrow in scope and therefore reveal little about whether GenAI, and LLMs specifically, possess and can deliver broader, conceptually grounded marketing knowledge

across marketing as a field. Strong performance on these tasks may signal some understanding of the underlying marketing principles, but it may also just reflect the ability to produce effective output while lacking deeper comprehension. One encouraging but isolated piece of evidence comes from Jürgensmeier & Skiera (2024), who, in the context of two marketing analytics exams, show that LLMs can provide feedback that resembles human graders. However, this evidence is limited in scope (to a single marketing analytics course) and thus leaves open the broader question of how much LLMs know about the general corpus of marketing knowledge, and whether they can serve as reliable tools for conveying this knowledge and guiding decision-making.

We contribute to this literature by shifting the focus from what LLMs can *do* in marketing to what they *know* about it and assessing the conceptual grounding of LLMs' interactions with marketing practitioners.

2.2 Knowledge-Based Systems in Marketing

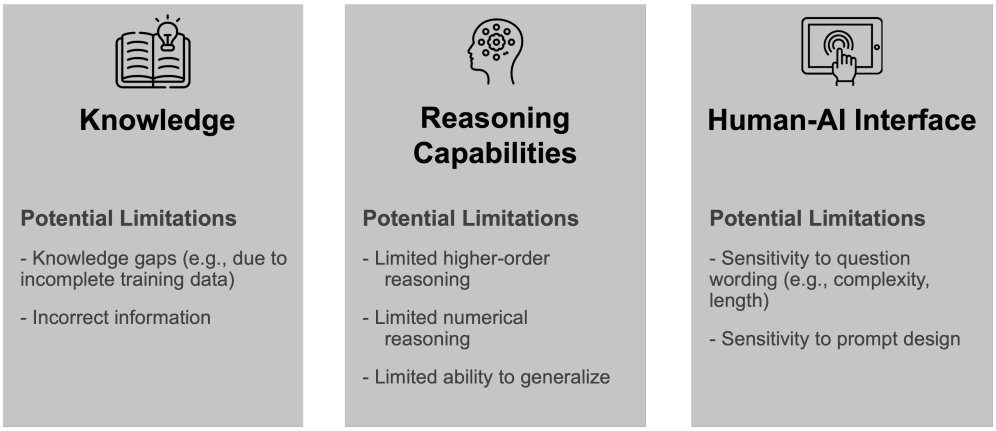
Long before modern AI, marketers developed expert systems to encode and convey specialized knowledge in areas like advertising, negotiations, and sales promotions (e.g., Keon & Bayer, 1986; Rangaswamy et al., 1989; Burke et al., 1990; Wierenga, 1990). These expert systems relied on manually constructed knowledge bases composed of explicit rules and facts (Wierenga & Van Bruggen, 1997). They reasoned via a transparent, logic-based inference engine that applied a sequence of rules to reach conclusions. While initial results were promising, their knowledge bases required substantial effort to develop and maintain (Burke, 1991), which hindered their wider adoption.

LLMs can serve similar functions, but rely on fundamentally different mechanisms. Rather than encoding knowledge explicitly, they embed information implicitly in their connectionist model parameters during training on large-scale text data. While this implicit representation eliminates the need for manually constructing knowledge bases, it introduces new challenges: the embedded knowledge is not directly observable and must be inferred through probing. Likewise, their reasoning capabilities emerge from learned statistical patterns rather than a separate inference engine, which

makes them challenging to audit. Moreover, LLMs are optimized for next-token prediction rather than truthful generation, and as their outputs involve stochastic sampling, their reliability is not structurally ensured.

Our study extends this literature by testing whether LLMs can function as a novel type of marketing knowledge system that bypasses the constraints of explicit knowledge representation required by earlier systems. To structure our evaluation, we adopt a framework from prior work on knowledge-based systems that defines three key components: (1) a knowledge base that stores information, (2) an inference engine that determines reasoning capabilities, and (3) a user interface that governs user interactions with the system (Rangaswamy et al., 1989). While LLMs do not mirror these components technically, we argue that they exhibit analogous features, with encoded knowledge, reasoning capabilities, and a human-computer interface. Thus, this analogy offers a useful framework for evaluating their effectiveness as marketing knowledge systems (see Figure 1).

Figure 1: Key Dimensions of Our Evaluation of LLMs as Knowledge Systems



Accordingly, we focus on three potential bottlenecks that may limit LLMs' effectiveness as marketing knowledge systems: (1) knowledge limitations, such as misconceptions or gaps; (2) reasoning limitations, such as strong recall but weak numerical reasoning; and (3) interface constraints, such as sensitivity to prompt design or question wording.

2.3 AI Benchmarking

As AI capabilities expand, benchmarking has become essential for evaluating and comparing systems. Methods vary by task: researchers have used annotations to evaluate LLMs' natural language understanding (e.g., ANLI, Nie et al., 2020), text continuation or Winograd schemes to assess common sense reasoning (Levesque et al., 2012; Zellers et al., 2019), and programming challenges to evaluate coding abilities (Chen et al., 2021). Knowledge-based benchmarks often use question-answer formats, including general datasets like TriviaQA (Joshi et al., 2017) or domain-specific ones like GPQA (Rein et al., 2024). More recent efforts push toward expert-level evaluation, such as Humanity's Last Exam (Phan et al., 2025), which tests frontier models on 2,500 questions across dozens of academic disciplines, or toward real-world task performance, such as GDPval (Patwardhan et al., 2025), which evaluates AI on professional deliverables across 44 occupations.

Yet, no benchmark to date comprehensively assesses LLMs' knowledge of marketing. Broader benchmarks, like MMLU (Hendrycks et al., 2021) or MMLU-Pro (Wang et al., 2024), provide only limited marketing coverage; for instance, MMLU includes just 264 marketing-related questions.[1] Our work contributes by introducing a novel, human-constructed benchmark specifically designed to evaluate the marketing knowledge of current and future LLMs. This resource can also be extended beyond marketing, serving as a blueprint to develop similar benchmarks and evaluation methods in other domains.

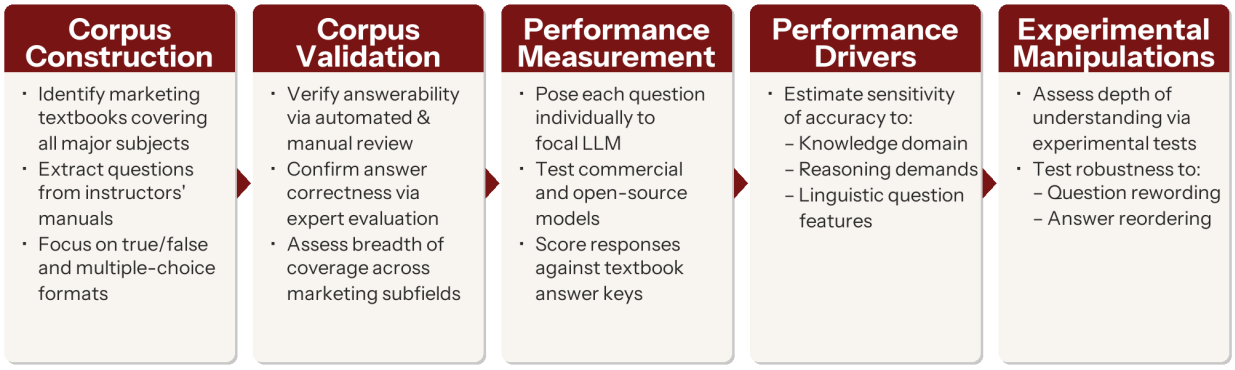
3 Data and Method

Our empirical study of LLMs' marketing knowledge entails the following steps, as illustrated in Figure 2. First, we construct a corpus of question-answer pairs derived from supplementary materials of academic marketing textbooks. Second, we validate the corpus along several quality dimensions. Third, we measure LLM performance by posing each question to multiple models and scoring their responses. Fourth, we identify the drivers of performance variation across knowledge domains, reasoning demands, and linguistic features. Fifth, we test the robustness of our findings

[1] See https://huggingface.co/datasets/cais/mmlu/viewer/marketing.

through experimental manipulations of the question texts. The following sections detail each step.

Figure 2: Our Research Approach



3.1 Corpus Construction

Our aim is to capture the generally established body of foundational marketing knowledge represented in academic teaching, rather than marketplace knowledge like firm- or industry-specific insights. To construct this knowledge base, we queried major publishers and identified academic textbooks widely used in business education at top US business schools. We restricted our selection to recent editions published no earlier than 2018 and further required that each textbook have a separate official instructor manual. These manuals are companion resources in which questions and answers are provided separately from the textbook itself and typically available only to verified educators.

This process resulted in 25 textbooks spanning a broad range of marketing subjects (e.g., marketing management, consumer behavior, digital marketing), with an average copyright year of 2022 (range: 2018–2025). These textbooks come from three of the four largest U.S. textbook publishers, plus one smaller academic publisher focusing on upper-division and graduate audiences.[2] From their instructor manuals, we extracted 33,446 multiple-choice and true/false questions, along with 3,781 open-ended discussion questions. Together, these questions form our initial marketing knowledge corpus.

[2]Collectively, these 25 titles account for approximately 60% of commercial textbook adoptions in marketing at AACSB-accredited business schools, estimated based on a review of available public sources.

3.2 Corpus Validation

Before using this corpus to evaluate LLMs, we validate it along three dimensions: answerability, answer correctness, and breadth of coverage.

Answerability. To ensure that all questions are sound and can be answered reliably within the available context, we applied a combination of manual review and automated quality checks. We first verified that the questions used in our analysis can be answered based solely on the information provided, without requiring additional context or resources. We assessed this through a combination of automated and manual checks. First, we used a large language model (GPT-4o, accessed via OpenAI's API) to evaluate whether each question included sufficient information for an independent answer or relied on missing elements such as external graphs or tables. We provide the specific prompt used for this evaluation in Web Appendix B. After the automated assessment, two expert human coders manually reviewed all candidate questions—539 multiple-choice and 1,630 true/false—and removed those that failed to meet our criteria. In total, 320 multiple-choice and 136 true/false questions were excluded from the initial set of 33,446 questions, yielding a total of 32,990 questions. Examples of disqualified questions include those tied to specific points in time (e.g., "Which country spends the most on mobile advertising?"), dependent on textbook-specific content (e.g., "Based on Table 2.1 in the text..."), or referencing non-text elements (e.g., "Use the figure below...").

Answer correctness. Beyond answerability, we also verified that the correct answers provided in the test banks are in fact correct. To this end, we drew a random sample of 300 questions and independently verified each answer.[3] Two of the authors reviewed the sampled questions and confirmed that the designated correct answers were indeed accurate. In cases where a question fell outside the authors' domain expertise, we consulted subject-matter experts in the relevant marketing subfield. All answer keys were found to be accurate.

[3]These are the same 300 questions we later use in our human benchmarking study (Section 5).

Breadth of coverage. Finally, we assessed whether our knowledge base provides sufficiently broad coverage of the marketing discipline. Following Hambrick & Chen (2008), who conceptualize academic fields as social movements[4], we first identified scholarly communities in marketing that form around a shared intellectual interest, as captured by the 18 substantively oriented Academic Special Interest Groups (SIGs) maintained by the American Marketing Association.[5] Using GPT-4o, we mapped each question to the SIG whose topical focus best matched the question's content. Table 1 presents the resulting distribution of questions across subfields. Our knowledge base covers all 18 marketing subfields, but coverage is non-uniform: core subfields such as Marketing Research and Consumer Behavior are more prevalent than more specialized (e.g., Entrepreneurial Marketing) or more recent (e.g., Artificial Intelligence) subfields. We report additional details on the classification methodology in Web Appendix C.

Table 1: Coverage of Marketing Subfields in Our Knowledge Base

AMA SIG Subfield	N Books	N Questions (MC)	N Questions (TF)	N Questions (Total)
Marketing Research	25	5,204	2,659	7,863
Consumer Behavior	25	3,739	1,463	5,202
Marketing Communications	23	3,100	1,590	4,690
Retail and Pricing	25	2,699	1,109	3,808
Marketing Strategy	25	2,154	953	3,107
Selling and Sales Management	23	1,087	425	1,512
Global Marketing	20	852	397	1,249
Marketing and Society	23	703	405	1,108
Branding	19	693	349	1,042
Service Marketing	24	738	250	988
Technology	25	529	276	805
Inter-organizational	17	252	151	403
Organizational Frontlines Research	23	247	118	365
Relationship Marketing	21	218	104	322
Sports and Sponsorship-Linked Marketing	15	111	86	197
Sustainable Marketing	17	134	61	195
Entrepreneurial Marketing	18	62	50	112
Artificial Intelligence	9	18	4	22

Notes: Each question is classified into the AMA SIG whose topical focus best matches its content using GPT-4o. N Books = number of textbooks in our corpus that contain at least one question classified into the respective subfield. MC = multiple-choice; TF = true/false.

[4]We thank an anonymous reviewer for pointing us to this reference.

[5]See https://www.ama.org/academic-special-interest-groups/. We excluded three SIGs that are not substantively oriented: Doctoral Students, Marketing for Higher Education, and Teaching and Learning.

3.3 Performance Measurement

To evaluate LLMs' marketing knowledge, our main analysis focuses on their performance based on all 32,990 multiple-choice and true/false questions, which allow for objective assessment at scale. We pose each question individually to each LLM via its API. All prompts follow a standard Role-Task-Format (RTF) structure and were not iteratively optimized; we used straightforward formulations on the first attempt.

To assess both temporal progress and cross-provider variation, our analysis spans a longitudinal comparison within the OpenAI model family (from GPT-3.5 Turbo to GPT-5.2) and cross-sectional comparisons across the major providers (Meta's Llama, Anthropic's Claude, Google's Gemini). We prompt each model individually through the respective APIs provided by OpenAI (https://platform.openai.com/docs/overview), Google (https://ai.google.dev), and Anthropic (https://www.anthropic.com/api). For Llama models, we access the model via Together AI's API (e.g., https://www.together.ai/models/llama-3-1-405b). We keep model parameters at their default values to approximate the experience of a typical user interacting with these models through their standard interfaces (e.g., web chat). Web Appendix D summarizes the exact model versions and API settings we used.

The default prompt instructs the model to act as a marketing expert ("*You are a professor of marketing with many years of experience.*") and to provide the correct answer without additional explanations. For true/false questions, we include the instruction, "*Answer the question at the end True or False only,*" while for multiple-choice questions, we specify, "*Answer the question at the end a, b, c, d [or e] only.*" We then append the full question text to the prompt. To prevent dependencies between questions, we submit each query in a new chat session so that prior interactions cannot influence subsequent responses. We also test alternative prompt variations to assess the sensitivity of our results to the specific way models are instructed (see Web Appendix E).

The focal outcome of interest is *Correct*, a binary indicator equal to 1 if the model's answer matches the correct answer from the test bank (for multiple-choice questions) or correctly identifies

a statement (for TF questions).[6]

3.4 Performance Drivers

To examine the drivers of LLM performance, we extract features from each question aligned with the three key dimensions of our evaluation framework: knowledge, reasoning, and human-AI interaction.

To evaluate LLMs' **knowledge** coverage—from common to specialized areas—we first embed all questions using a transformer model, apply dimensionality reduction, and then cluster them based on their embedding distances.[7] Each cluster is labeled with a topic name (e.g., Market Research, Competitive Strategy) based on the common theme among the questions it contains. For each question i, we compute its Euclidean distance d_i to its topic centroid and define *Centrality* as $c_i = 1 - \frac{d_i}{d_{\max}^{(k_i)}}$, where $d_{\max}^{(k_i)}$ is the maximum distance within cluster k_i (see Web Appendix F for further details).

To evaluate LLMs' **reasoning abilities**, we draw on Bloom's revised taxonomy (Krathwohl, 2002), which classifies cognitive demands along two axes: the type of knowledge involved and the complexity of the cognitive process required. Specifically, we classify each question along two dimensions. The first is *Knowledge Type*, which captures what kind of knowledge is needed to answer a question: *Factual* knowledge involves basic terminology, definitions, and specific details (e.g., "What is a SKU?"); *Conceptual* knowledge concerns the relationships among ideas, such as principles, models, and theories (e.g., "How does brand equity relate to customer loyalty?"); *Procedural* knowledge pertains to methods, techniques, and knowing how to carry out a task (e.g., "How do you calculate customer lifetime value?"); and *Metacognitive* knowledge involves awareness of one's own reasoning processes and when to apply particular strategies.

[6] Because some models can be verbose, exact matching on the first characters of the response was not feasible. Instead, we use regular expression matching to locate the first occurrence of a choice option followed by a punctuation sign (any of ")", ".", ",", "-", or ":"). For example, we identify the response "The correct answer is A) e-WOM." as correct (when the textbook answer is A) because it matches "A)".

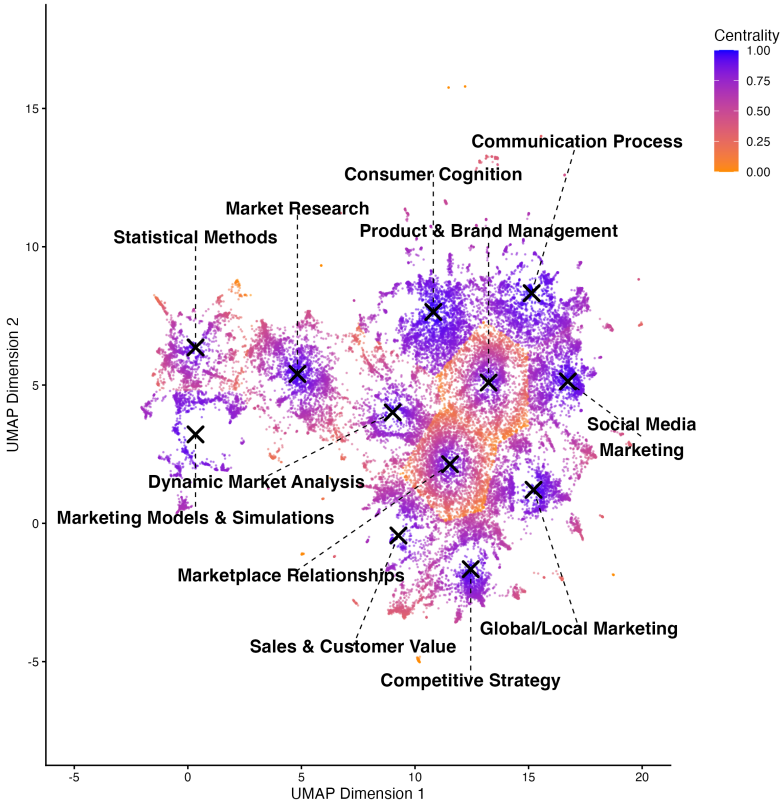
[7] While we used the AMA SIG-based classification in Table 1 to verify the breadth of coverage of our knowledge base, we adopt this data-driven clustering approach to identify topics for our analysis because it allows us to quantify both topic membership *and* centrality within a topic through the continuous embedding vectors. In addition, the AMA SIG-based classification exhibits highly uneven coverage, with several subfields containing very few observations (e.g., Artificial Intelligence: 22 questions), which limits their usefulness as regressors.

The second dimension is *Cognitive Process*, which captures the complexity of the mental operation required: retrieving relevant knowledge from long-term memory (*Remember*), comprehending the meaning of information (*Understand*), carrying out or using a procedure in a given situation (*Apply*), breaking material into its constituent parts and detecting how they relate to one another (*Analyze*), making judgments based on criteria and standards (*Evaluate*), and putting elements together to form a novel, coherent whole (*Create*). For classification, we use GPT-4o prompted with definitions from the taxonomy and instructed to label each question accordingly, a classification approach increasingly adopted in social science research (e.g., Rathje et al., 2024). We also code whether a question involves *Numerical Reasoning* (1 if the question text includes at least three numbers; 0 otherwise).

To evaluate LLMs' sensitivity to the **human-AI interface**, we focus on question wording and measure *Question Length* as the number of characters in the question text, and *Lexical Sophistication*, measured as the proportion of rare words in each question; that is, words not found in the 10,000 most frequent words of the Brown corpus or standard stopword lists (Kučera & Francis, 1967). Full classification details are provided in Web Appendix G.

Data overview. Our main dataset comprises 22,540 multiple-choice (MC) and 10,450 true/false (TF) questions spanning 12 marketing topics (= question clusters). Figure 3 visualizes our question base, including the underlying marketing topics, while sample questions are shown in Appendix A. Table 2 summarizes the distribution of question features used to evaluate potential limitations in LLMs' knowledge, reasoning abilities, and/or user interactions.

Figure 3: UMAP Visualization of our Knowledge Base



Notes: Each point represents a question from our knowledge base (N = 32,990). Questions are embedded using a transformer model and projected to two dimensions via UMAP (McInnes et al., 2018), such that similar questions appear nearby. Each X marks the mean UMAP vector of a topic identified via K-Means clustering. Topic labels are manually assigned based on shared themes. Colors indicate within-topic centrality, i.e., proximity to the topic centroid.

Table 2: Descriptive Statistics

Variable	Min	Mean (SD)	Max	Categories (%)
Marketing Topic	–	–	–	Marketplace Relationships (11.0%); Communication Process (8.3%); Competitive Strategy (8.4%); Consumer Cognition (10.9%); Dynamic Market Analysis (7.3%); Global/Local Marketing (7.8%); Market Research (10.4%); Marketing Models & Simulations (5.0%); Product & Brand Management (9.6%); Sales & Customer Value (6.2%); Social Media Marketing (8.7%); Statistical Methods (6.6%)
Centrality	0	0.64 (0.21)	1	–
Knowledge Type	–	–	–	Factual Knowledge (61.7%); Conceptual Knowledge (34.3%); Procedural Knowledge (3.9%); Metacognitive Knowledge (0.1%)
Cognitive Process	–	–	–	Remember (54.8%); Understand (19.6%); Apply (2.1%); Analyze (6.8%); Evaluate (16.6%); Create (0.04%)
Numerical Reasoning	0	0.03 (0.18)	1	–
Question Type	–	–	–	Multiple (68.3%); TF (31.7%)
Question Length	18	243.54 (171.76)	3,194	–
Lexical Sophistication	0	0.11 (0.08)	0.7	–

Note: We refrain from interpreting the "Create" and "Metacognitive Knowledge" levels in what follows as they remain basically absent in our knowledge base.

Estimation approach. To explore the sensitivity of LLM performance along these question characteristics, we estimate logistic regressions with *Correct* as our dependent variable.[8] An obvious concern is the potential confounding effect if some textbooks being tested were used as part of the LLM training data, such that performance differences might not necessarily reflect variation in the models' general marketing knowledge but rather whether certain textbooks—and possibly the corresponding questions—were included in their training corpus. Likewise, different authors may take different approaches to writing questions, such that question design and difficulty may vary systematically across textbooks. To address these challenges, we include textbook-level fixed effects. Our intuition is that if a textbook appeared in the model's training data, the entire material is likely included, driving performance differences *across* textbooks. Residual variation within a single textbook, instead, is more likely to reflect how the model's performance varies more generally, e.g., across knowledge types or marketing topics. Likewise, this approach allows us to control for potential author-specific differences (e.g., in difficulty). We estimate separate regressions for each tested model and use the resulting coefficients to compute average marginal effects (AMEs) along each dimension.

3.5 Experimental Manipulations

The preceding steps identify to what degree LLM accuracy varies across question characteristics. However, these observational patterns alone do not reveal the extent to which baseline performance might be driven by LLMs recognizing specific questions encountered during training rather than from genuine marketing understanding. To address this question, we implement two controlled manipulations across our question set: (1) *Rewording*, where we change a question's wording without altering its meaning; and (2) *Shuffling Multiple-Choice Answers*, where the question remains unchanged, but answer choices are reordered. Our intuition is that if the model's performance relies heavily on strict pattern recognition, even slight linguistic changes should lower accuracy. Similarly, if performance is driven by memorized answer keys, often presented as "Correct Answer: C" in solution manuals, then shuffling answer options should impair the model's ability to choose

[8] Results remain consistent for alternative model specifications, like a linear probability model.

correctly. Further details and examples of both manipulations are provided in Web Appendix H.

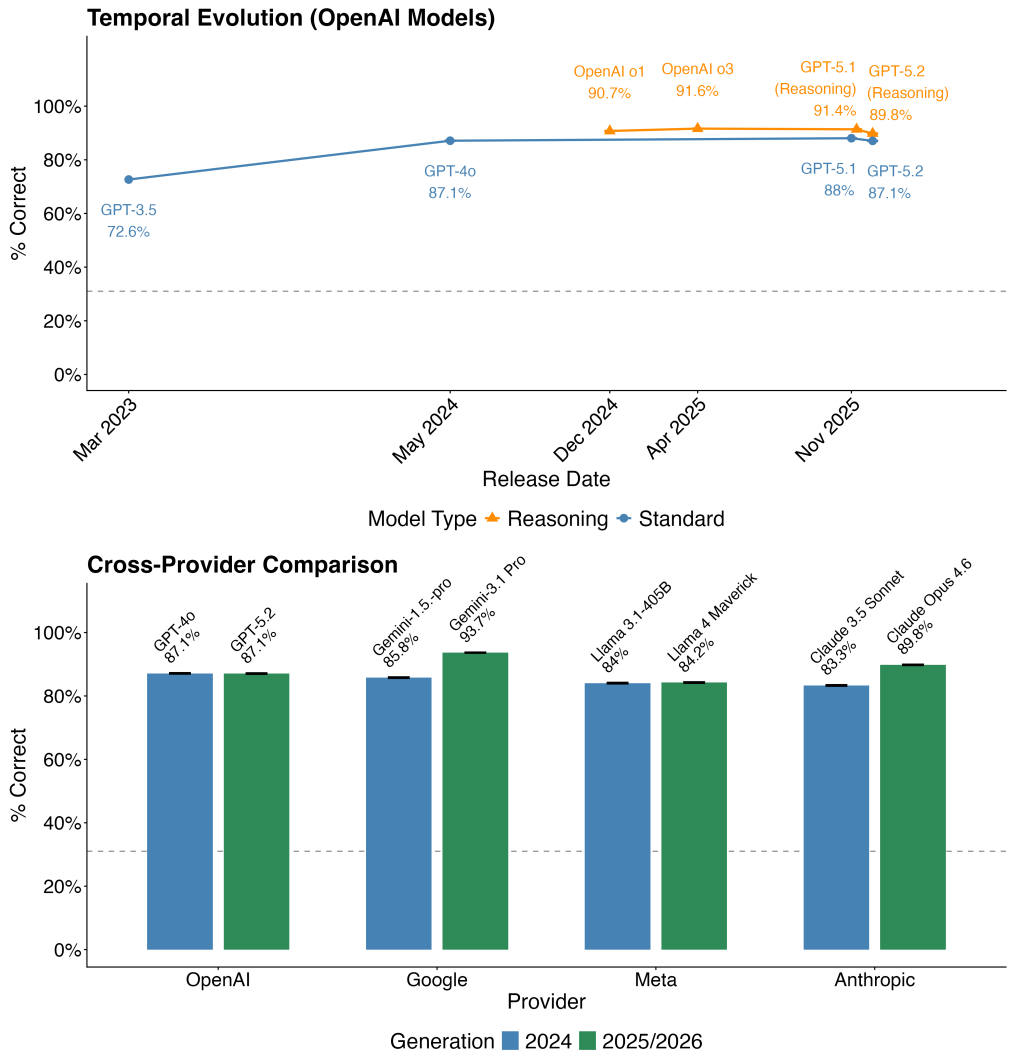
4 Empirical Results

4.1 Descriptive Results

We begin by reporting LLMs' baseline accuracy in answering marketing questions from our knowledge base. Figure 4 summarizes our results. The top graph shows the temporal evolution of accuracy within the OpenAI model family: from GPT-3.5 (72.6%) to GPT-4o (87.1%) and GPT-5.1/5.2 ($\sim$87–88%), we observe a substantial increase in accuracy over time. Most recent increases are driven by reasoning models (OpenAI o1, o3, and GPT-5.1/5.2 Reasoning), which push accuracy up to 91.6%. The bottom graph compares accuracy of the OpenAI model family to other providers' models (Google, Anthropic, Meta) in 2024 and towards 2025/2026. Performance is high across all models, with the proportion of correct answers ranging from $\sim$84.2% to $\sim$93.7% as of 2025/2026. Google's Gemini shows the most notable increase in accuracy compared to the preceding generation (from 85.8% to 93.7%) and emerges as the top-performing model overall.

We next explore how this performance varies across the three dimensions knowledge, reasoning abilities, and human-AI interaction using the estimation approach described in Section 3.4 and models from the OpenAI family.

Figure 4: Baseline Accuracy Across Models



Notes: The dashed horizontal line indicates the expected share of accurate answers under random guessing across the entire question database (31%).

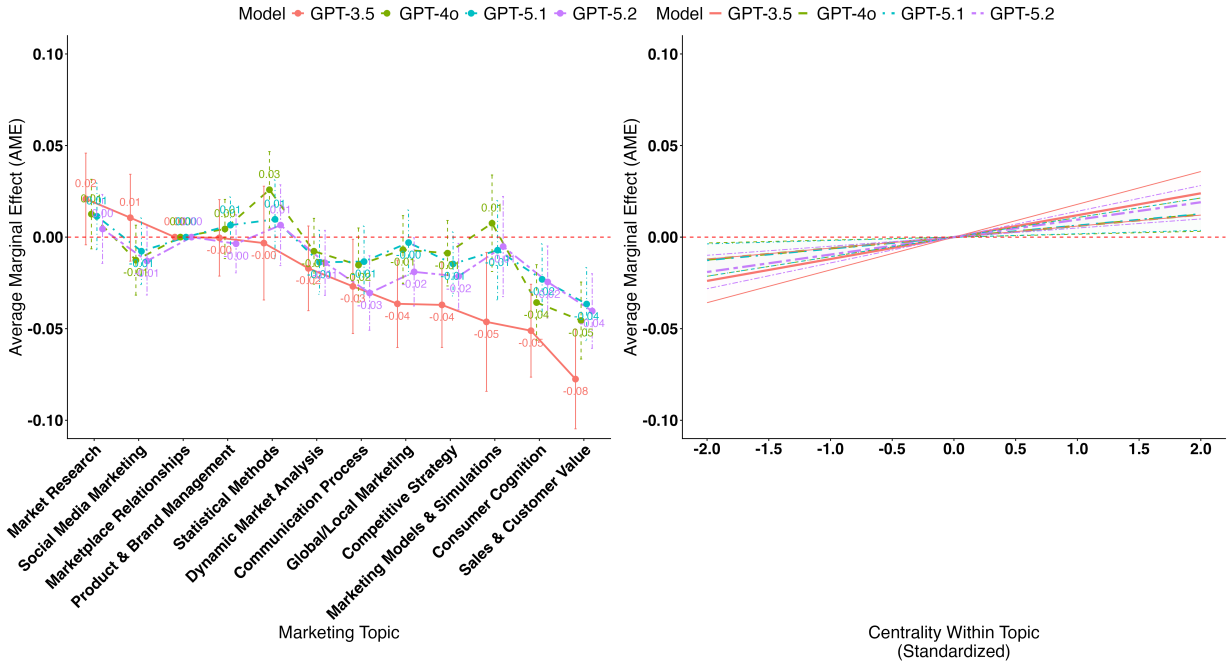
4.2 Sensitivity Analysis

In the following, we present Average Marginal Effects (AME) plots across the three dimensions for the main GPT model generations (GPT-3.5, GPT-4o, GPT-5). The underlying regression results are provided in Web Appendix I.

We begin with the **knowledge dimension** by focusing on *Marketing Topics* and *Centrality* (Figure 5). While GPT-3.5 shows moderate sensitivity (e.g., +/- 5% across topics), after GPT-4o, both variables are only modestly associated with the likelihood of a correct answer. Most marketing

topics show estimated marginal effects close to zero; only a few deviate by approximately +1 to –4 percentage points for GPT-5.1/5.2. For *Centrality*, average marginal effects are small across all models (as low as ∼0.6% per standard deviation even for the outdated GPT-4o): questions that are more central to a topic are only slightly easier to answer.

Figure 5: Marginal Effects Across the Knowledge Dimension



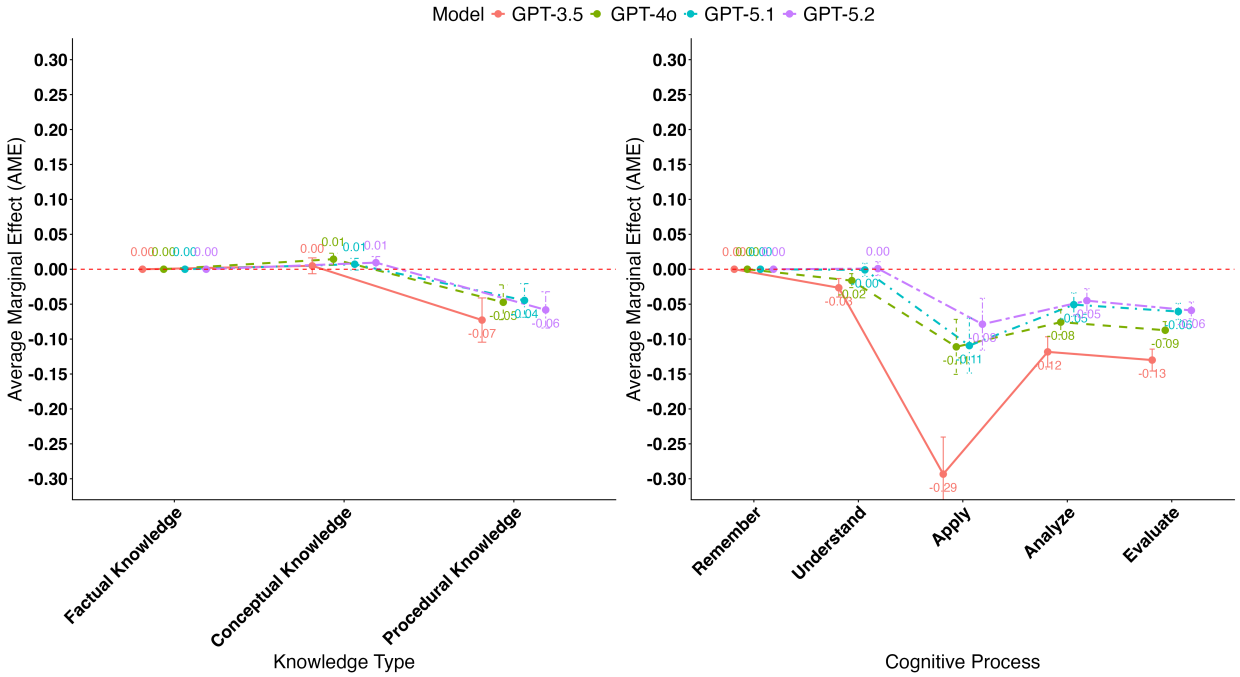
Notes: Average marginal effects based on the full model specification (see Web Appendix I for full regression results), estimated separately for each model. Marketplace Relationships is the reference category for Marketing Topic. Error bars/shaded regions indicate the 95% confidence interval.

Second, we examine the **reasoning dimension**: *Knowledge Type* and *Cognitive Process* (Figure 6). Both show a substantive relationship with answer accuracy. Performance remains high when processing factual or conceptual knowledge across all model generations, but declines moderately for procedural knowledge (i.e., methods or frameworks).[9] For cognitive processes, all models, including the early GPT-3.5, show comparably strong recall and understanding, but accuracy decreases for higher-order reasoning tasks such as applying, analyzing, or evaluating. This pattern shows clear temporal improvement: while GPT-4o exhibits accuracy declines of ∼8-11% for these

[9]We refrain from interpreting the results for 'Metacognitive Knowledge' and the cognitive process 'Create' due to their very low number of observations.

higher-order tasks, GPT-5.2 narrows these gaps to 5-8%.

Figure 6: Marginal Effects Across the Reasoning Dimension

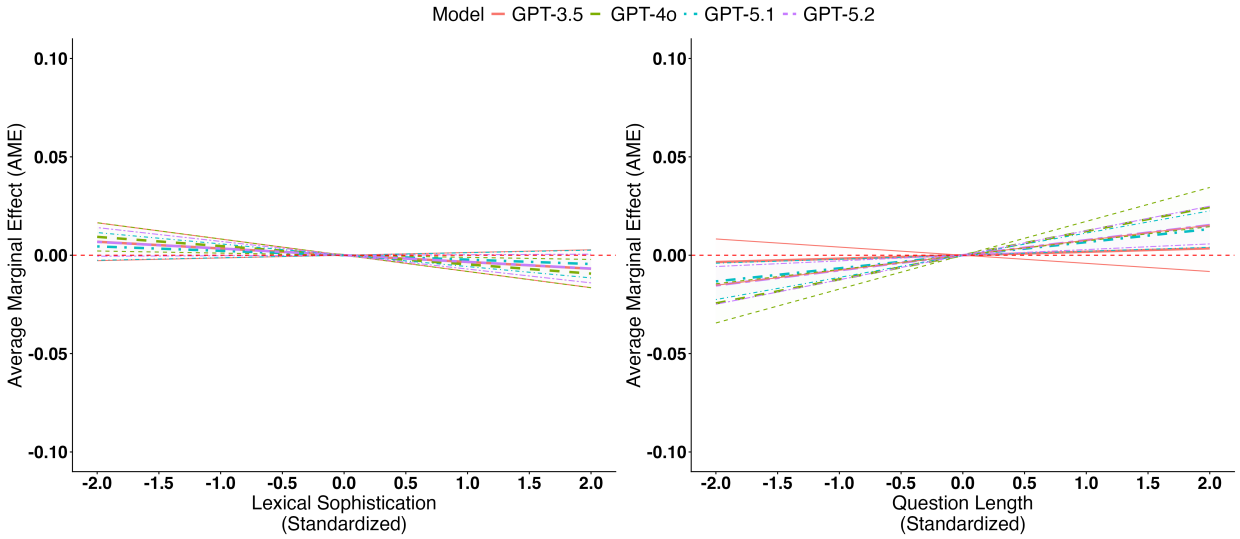


Notes: Average marginal effects based on the full model specification (see Web Appendix I for full regression results), estimated separately for each model. Reference categories are Factual Knowledge (Knowledge Type) and Remember (Cognitive Process). Error bars indicate the 95% confidence interval.

Third, we examine the **human-AI interaction dimension** (Figure 7). Across all model generations, accuracy is largely unaffected by features related to user-AI interactions. *Lexical Sophistication* has minimal impact, while longer questions—such as those providing more context—have a modest positive effect. Prompt design variations similarly show little effect on performance (∼-0.90%, see Web Appendix E).

In summary, our sensitivity analysis reveals that LLMs possess a broad marketing knowledge base that spans the entire spectrum of topics, including niche areas. Knowledge coverage improved across model generations. Reasoning abilities show sensitivity to the cognitive process dimension, particularly once questions move beyond remembering and understanding, though newer models show clear gains. In contrast, the human-AI interaction dimension shows minimal sensitivity to question wording.

Figure 7: Marginal Effects Across the Human-AI Interaction Dimension

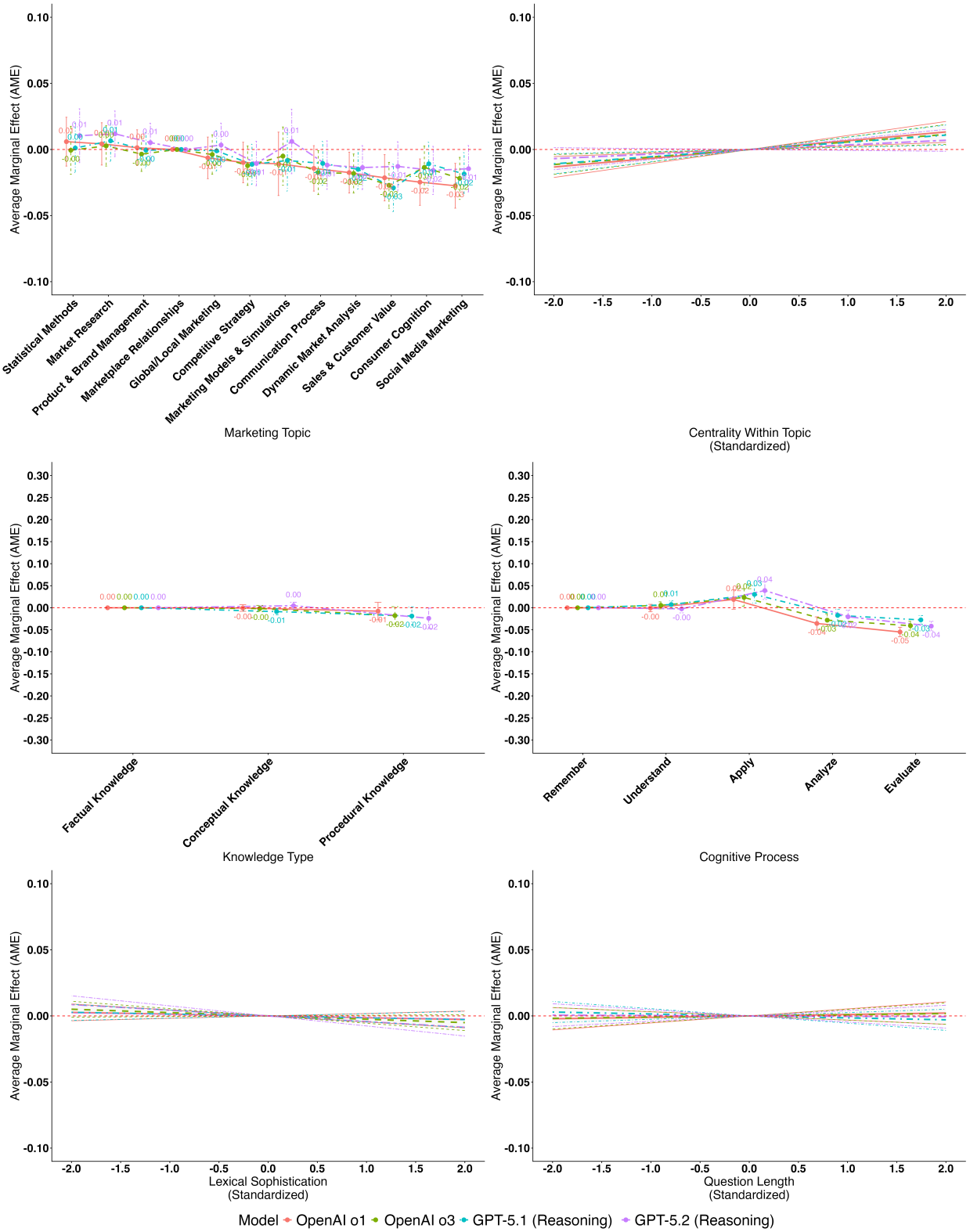


Notes: Average marginal effects based on the full model specification (see Web Appendix I for full regression results), estimated separately for each model. Shaded regions indicate the 95% confidence interval.

Reasoning models. A natural question is whether dedicated "reasoning models," which employ extended inference-time computation, can further close the remaining gaps in higher-order reasoning. Figure 8 presents AMEs for four reasoning-focused models (OpenAI o1, o3, and GPT-5.1/5.2 Reasoning).[10] Indeed, reasoning models substantially reduce the accuracy declines associated with applying, analyzing, and evaluating, with application being the highest-performing cognitive process. Declines are also much less pronounced for 'analyze' and 'evaluate' questions. Meanwhile, both the knowledge and human-AI interaction dimensions remain stable.

[10]GPT-5.1 and GPT-5.2 offer configurable reasoning effort. For the reasoning variants presented here, we set the `reasoning.effort` parameter to `high` (OpenAI, 2025).

Figure 8: Reasoning Models: Marginal Effects Across All Dimensions



Notes: Average marginal effects based on the full model specification (see Web Appendix I for full regression results), estimated separately for each reasoning model. Error bars/shaded regions indicate the 95% confidence interval.

Numerical reasoning and true/false asymmetry. Two additional patterns in our data merit closer examination across model generations. First, questions involving numerical reasoning are consistently harder for LLMs. Table 3 reports the corresponding AMEs: GPT-3.5 shows an accuracy decline of approximately -20% on numerical questions, which narrows to roughly -10% for GPT-4o and GPT-5, and to -2% for reasoning models. Second, for true/false questions, all standard models are substantially more accurate when confirming true statements than when identifying false ones. One interpretation is that LLMs, trained to be helpful and to some degree agreeable, may be less inclined to reject a user-presented statement. From a different perspective, findings in cognitive psychology show that rejecting false propositions requires more cognitive effort than affirming true ones (e.g., Gilbert, 1991). The gap is large for GPT-3.5 (+25%) and GPT-4o (+21%), remains at a substantial +11% for GPT-5.1/5.2, and is only reduced further for reasoning models. Table 3 summarizes these trends across model generations.

Table 3: Numerical Reasoning and True-False Asymmetry Across Models

Panel A: Standard Models				
	GPT-3.5	GPT-4o	GPT-5.1	GPT-5.2
Num. Reasoning	-19.96%***	-10.58%***	-9.75%***	-8.78%***
	(1.52%)	(1.01%)	(0.97%)	(1.05%)
T/F Asymmetry	25.19%***	21.31%***	11.35%***	11.06%***
	(0.93%)	(0.80%)	(0.77%)	(0.79%)
Panel B: Reasoning Models				
	OpenAI o1	OpenAI o3	GPT-5.1 (Reasoning)	GPT-5.2 (Reasoning)
Num. Reasoning	-2.56%**	-2.50%**	-2.35%**	-2.15%*
	(1.05%)	(0.99%)	(1.01%)	(1.15%)
T/F Asymmetry	7.11%***	3.30%***	-0.89%	4.62%***
	(0.71%)	(0.68%)	(0.67%)	(0.73%)

Notes: Numerical Reasoning shows the AME of the numerical reasoning indicator on the probability of a correct answer. T/F Asymmetry reports the accuracy gain, in percentage points, for true/false questions when the textbook answer is "True" relative to when it is "False." Standard errors in parentheses. Significance levels: $^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$.

4.3 Manipulation Results

We next report how LLM performance responds to the controlled manipulations described in Section 3.5.

4.3.1 Overall Manipulation Effects

Table 4 presents the impact of both manipulations (question rewording & shuffling answer options) on the likelihood of answering a question correctly. The effects are modest: rewording reduces the likelihood of a correct answer by approximately 2% while shuffling answer options lowers it by about 5% for multiple-choice questions. These results suggest that some questions may be answered through pattern matching or memorization, but offer no strong evidence that these mechanisms are the primary basis for LLMs' overall strong performance.

Table 4: Manipulation Results: Average Marginal Effects on Probability of Correct Answer

Manipulation	AME	95% CI Lower	95% CI Upper	N
Reword	−0.022	−0.027	−0.017	32,990
Shuffle MC	−0.047	−0.053	−0.040	22,540

Notes: Average marginal effects from a logistic regression with *Correct* as the dependent variable and *treated* as the independent variable. For 'Reword', *treated* = 1 for 32,990 reworded questions; for 'Shuffle MC', *treated* = 1 for 22,540 multiple-choice questions with shuffled answer options.

4.3.2 Textbook-Level Contamination Analysis

While the overall shuffling effect is modest, it raises the question of whether some textbooks in our corpus may be affected by potential training data contamination. Although our main analysis already controls for textbook fixed effects, there remains the risk that only some questions may have been available in the LLMs' training data—for example, through leaked materials or content adopted by instructors in publicly available assignments or exams. If that is the case, the shuffling effect should reveal such contamination: we would expect answer-order memorization to be concentrated in those textbooks, while others should remain unaffected. We can therefore use the shuffling results to identify and remove questions from potentially problematic textbooks and test the robustness of our findings.

Specifically, we estimate the shuffling effect separately for each of our 25 textbooks using a linear probability model with *Correct* as the dependent variable and a binary indicator for whether answer options were shuffled as the independent variable. Focusing on GPT-4o, we identify textbooks where shuffling significantly reduces accuracy ($p < .10$).[11] This procedure flags 13 out of 25 textbooks as potentially affected by answer-order memorization, while the remaining 12 textbooks show no significant shuffling effect.

We then re-estimate our main analysis excluding the 13 potentially affected textbooks, retaining only the 12 "clean" textbooks for which we find no evidence of memorization. For GPT-4o, average accuracy on this reduced sample is 87.7%, compared to 87.1% on the full sample. Likewise, regression results based on the clean subsample are substantively unchanged.

This result reinforces our conclusion that LLMs' strong performance is not primarily driven by memorized answer patterns. Even when we restrict the analysis to textbooks where we find no evidence of answer-order memorization, and where performance must therefore rely on the model's broader conceptual understanding, accuracy remains equally high. Full details, including per-textbook treatment effect estimates and a robustness comparison across GPT-family models, are provided in Web Appendix J.

4.4 Assessing Open-Ended Questions

Our main analysis focuses on multiple-choice and true/false questions, which allow for objective, scalable evaluation. A natural concern is whether the strong performance we observe is driven by the simplistic nature of these closed-ended question formats. To test this, we conducted a supplementary analysis using 3,781 open-ended discussion questions extracted from the same instructor materials, analyzing the performance of GPT-4o and its competitors from the same model generation.

[11] Since our goal is to exclude any textbook with even suggestive evidence of memorization, we deliberately use a lenient significance threshold and do not adjust for multiple testing. This errs on the side of over-exclusion and makes the subsequent robustness check more conservative.

4.4.1 LLM-Based Grading

As with the multiple-choice and true/false items, we posed each discussion question individually to each LLM. We adapted the prompt to fit the discussion format and instructed all models to keep their answers to about 50 words—roughly the average length of the textbook responses—to account for variations in response length across models. The prompt read:

> *You are a professor of marketing with many years of experience. Answer the following discussion question. Limit your answer to about 50 words.*

Once the LLM responses were collected, we asked a separate instance of GPT-4o to grade them against the textbook answers on a scale from 0 to 100, with 100 representing a fully correct response. We emphasized that high scores should reflect conceptual alignment with the textbook answer, not mere verbal similarity. The evaluation prompt is provided in Web Appendix K.

Table 5 reports the average scores for models from the GPT-4o generation across all discussion questions. All tested models perform well, with average scores ranging from 78.3 to 82.9—only slightly below their corresponding mean accuracies from multiple-choice and true/false items (see Figure 4). Model differences remain small (range: $\sim$4.6%), with GPT-4o ranking second behind Claude 3.5. While this grading approach has limitations, which we turn to next, the results offer reassurance that our main findings are not simply an artifact of the multiple-choice or true/false format. The consistency across an entirely different question type suggests that the models' performance reflects broader capabilities, not just proficiency with structured answer formats.

Table 5: Average Scores of LLM Answers to Discussion Questions

	Answer Score		
Model	Mean	SE	N
Claude 3.5 Sonnet	82.9	0.30	3,781
GPT-4o	81.7	0.27	3,781
Meta Llama 3.1-405B	80.7	0.31	3,781
Gemini-1.5-Pro	78.3	0.32	3,781

Notes: Answer Score is measured on a scale from 0 (fully incorrect) to 100 (fully correct).

4.4.2 Validation with Human Coders

A potential concern with using an LLM to grade LLM-generated responses is self-preferencing bias: LLMs could systematically rate their own or other LLMs' outputs more favorably than warranted, even without explicit knowledge of these answers' LLM-authorship (Chen et al., 2025). To address this, we randomly selected 50 discussion questions and recruited four independent human coders (PhD students in marketing, none of whom were members of the author team) to evaluate GPT-4o's answers against the textbook answers on the same 0–100 scoring rubric. Coders were blind to the LLM-assigned scores, to the origin of the answers, and to the objective of the study.

On this subsample, the LLM evaluator's mean score (79.2, SD = 19.9) is close to the human average (76.1, SD = 23.9), and the correlation between LLM scores and the average human score is high at 0.844. Full details and the complete correlation matrix are reported in Web Appendix K. These results suggest that our LLM-based grading approach provides a reliable assessment of open-ended response quality without systematic bias.

5 Human Benchmarking

5.1 Study Design

Our analysis so far characterizes the structural patterns of AI performance on marketing knowledge tasks. A natural question is how this performance compares to that of humans. While peak human performance is reasonably well understood—the very best students in carefully prepared test situations can perform at the 95% level or above on questions like ours—an open question is how much of that knowledge is actually available outside these highly prepared, artificial environments. We therefore evaluate the performance of human respondents across various proficiency levels and contrast it to our AI-based results.

We conducted multiple studies across three respondent groups: (1) undergraduate marketing students, (2) marketing professionals (MBA students), and (3) doctoral (PhD) students in marketing. We randomly selected 300 questions from our full corpus, balanced by cognitive process level to

ensure participants were exposed to the full range of question types (since simple "Remember" questions are overrepresented in the raw data). From this pool, each respondent answered 15 randomly drawn questions. In total, 323 participants provided 4,845 responses. Rather than relying on third-party market research companies (e.g., online panels) to recruit participants, we leveraged the student and alumni pool of a large US business school, which allowed us to verify that all participants genuinely belong to the claimed respondent groups. The studies were administered online, with various guardrails in place to prevent the use of external resources. In addition, we tested a subgroup of undergraduate respondents in a fully controlled lab setting to ensure we record their actual performance absent any technological aids; in that condition, participants were also incentivized for accuracy through a performance-based lottery. Full details on sampling, administration, and the subsample results across study conditions are provided in Web Appendix L.

5.2 Study Results

Table 6 reports overall accuracy by respondent group. Performance increases with expertise: undergraduates average 55.0%, MBA professionals 60.7%, and PhD students 63.0%. All groups fall well below the accuracy levels of current LLMs (83–94%).

Table 6: Human Benchmarking: Overall Accuracy by Respondent Group

Respondent Group	Mean	SE	N
Overall	56.0%	0.01	4,845
Undergraduate Students	55.0%	0.01	4,065
Marketing Professionals (MBA)	60.7%	0.02	615
PhD Students	63.0%	0.04	165

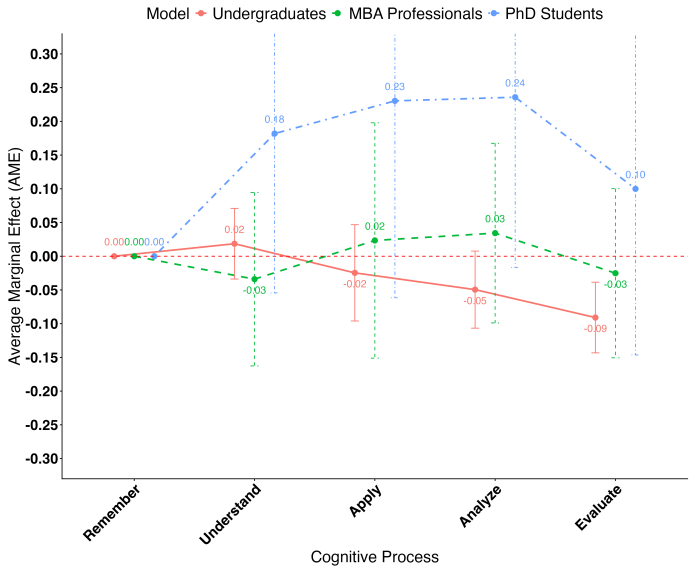
Notes: Accuracy reflects the proportion of correctly answered questions. N = number of question-level responses. Data collected across six studies with 323 participants total. Respondents answered a random subset of questions without prior preparation.

Beyond overall accuracy, Figure 9 examines how human performance varies across cognitive process levels for each respondent group. The patterns reveal a suggestive complementarity with AI performance.[12] As respondents advance in their marketing training, their relative strengths

[12]We acknowledge that these subsample analyses have limited statistical power and the subgroup estimates entail substantial uncertainty; we therefore treat these as suggestive rather than definitive evidence.

shift toward higher-order reasoning: MBA professionals show positive marginal effects for Apply (+2.3%) and Analyze (+3.4%), while PhD students show large positive effects across Apply (+23.0%), Analyze (+23.6%), and Evaluate (+10.0%), all relative to Remember. Thus, the more specialized human experts become, the more their performance profile diverges from that of LLMs, suggesting that the two are more complementary than substitutable when it comes to applying marketing knowledge. Pooled AME estimates and the full regression specification are reported in Web Appendix L.

Figure 9: Average Marginal Effects on Human Accuracy by Respondent Group



Notes: Baseline category: Remember. Effects estimated from separate logistic regressions per group with textbook fixed effects and question-level controls.

6 Discussion

6.1 Summary

Across a corpus of more than 30,000 questions spanning the established body of marketing knowledge, we find that LLMs have made rapid and substantial progress in encoding foundational marketing knowledge. Accuracy has risen from 72.6% (GPT-3.5) to 83–94% across current models from major providers.

Returning to the question that motivated this study—*can LLMs serve as marketing knowledge systems?*—we assess their capabilities along three dimensions and find that the answer is a qualified yes. First, the models exhibit near-complete knowledge coverage, with no systematic gaps across marketing topics. Second, they recall and understand facts and concepts with near-perfect accuracy. Performance declines moderately for higher-order reasoning tasks such as applying, analyzing, and evaluating, but newer model generations are closing these gaps, and reasoning models close them further still. Third, accuracy is stable across variations in question wording, linguistic complexity, and prompt design. Several lines of evidence indicate that this performance reflects genuine conceptual understanding rather than rote memorization. Our experimental manipulations (rewording questions and shuffling answer options) reduce accuracy only slightly (-2% to -5%); restricting the analysis to textbooks with no evidence of answer-order memorization leaves results unchanged and performance generalizes to open-ended discussion questions, where LLMs achieve average scores of 78–83 out of 100, validated by independent human coders. A human benchmarking study further confirms that LLMs substantially outperform human respondents at every level of training, from undergraduates (55%) to MBA professionals (61%) and PhD students (63%).

These findings indicate that a meaningful threshold has been crossed. Most, if not all, of the explicit marketing knowledge accumulated in academic textbooks is now broadly accessible through a conversational interface. In effect, the barriers to accessing foundational marketing knowledge have been substantially reduced. While access to knowledge alone is never sufficient for informed decision-making, this shift has far-reaching implications for marketing researchers, practitioners, and educators, which we turn to next.

6.2 Implications for Marketing Scholars

For academic researchers, who play a key role in developing, curating, and validating foundational marketing knowledge, our findings open new opportunities and challenges on the dissemination and practical impact of scholarly work.

Dissemination. While the potential of AI for *generating* marketing knowledge has received

considerable attention (e.g., assisting in designing experiments, generating stimuli, or replacing participants; Blanchard et al., 2025; Garvey & Blanchard, 2025; Toubia et al., 2025), less focus has been placed on how these same models might affect the *dissemination* of academic knowledge into practice. Our results suggest that this may become an important role for LLMs, which can serve as knowledge systems through which foundational insights can enter managerial decision-making. Textbooks have long functioned as a primary vehicle through which research is synthesized and passed on to practitioners and students. Much of the knowledge in our corpus originated as academic research that was later incorporated into teaching materials. To the extent that LLMs accurately encode this textbook content, as suggested by our findings, they are already serving as a downstream dissemination channel for decades of marketing scholarship. LLMs could thus significantly expand the reach and influence of academic research. Historically, even highly relevant research has not always realized its full potential impact because practitioners lack the time or means to discover and apply it. LLMs can reduce such frictions.

Control. A notable observation from our study is that the models we evaluate are general-purpose LLMs, none of which were fine-tuned or otherwise specialized for marketing. Yet they already encode a highly accurate representation of foundational marketing knowledge. On the one hand, this is encouraging: any practitioner, student, or researcher using a commercially available LLM already has access to reliable marketing knowledge without requiring domain-specific customization. On the other hand, it implies that the marketing discipline currently has comparatively little control over how its knowledge is represented and disseminated through these systems. For the well-established, foundational knowledge tested in our study, this may be unproblematic. However, it becomes more consequential when considering the more fluid frontier of current research, presented in journal articles, or knowledge that requires careful interpretation and contextualization. In such cases, how an LLM synthesizes and communicates insights matters, and that process is currently governed entirely by the AI companies that build and train these models.

Discipline-owned systems. This points to a potential opportunity for the marketing discipline to develop its own AI-based knowledge systems. One could envision, for instance, building on a

capable open-source foundation model (such as a Llama model, which performed reasonably well in our evaluation) that already encodes the general marketing knowledge base. This foundation could then be extended through post-training on a curated corpus of high-quality academic content to keep the knowledge base current, and/or through fine-tuning on collectively curated conversations between marketing experts and users to shape how that knowledge is communicated, much like how AI companies currently fine-tune their general models on proprietary dialogue corpora. Such an approach could give the field its own interface to the collective marketing knowledge base, one where the discipline retains agency over how established and emerging knowledge is represented, contextualized, and conveyed to relevant stakeholders.

6.3 Implications for Marketing Managers

For marketing practice, our findings carry implications at the individual, organizational, and market level.

Individual implications. On the individual level, our results suggest that LLMs can provide instant access to foundational marketing knowledge. They can help managers acquire or refresh relevant concepts, map decision contexts onto established frameworks (such as the 5 Cs, STP, or Porter's Five Forces), and surface implicit assumptions that shape how problems are framed. A product manager can evaluate a pricing decision against established frameworks, a brand manager can draw on consumer behavior theory to interpret surprising research findings, and a startup founder without formal training can develop a segmentation strategy grounded in established principles. Such guidance is more robust than relying solely on a manager's current mental model and more accessible than consulting textbooks. Our human benchmarking results suggest a natural complementarity: as professionals advance in their training, they develop stronger higher-order reasoning. LLMs, with near-perfect performance on lower-order tasks, can provide the conceptual scaffolding while the marketer contributes contextual judgment and critical evaluation.

Of course, not all marketing problems are resolvable through foundational knowledge alone (Wierenga, 2002), and many decisions require analogizing, intuition, or creativity (Wierenga &

Van Bruggen, 1997). Still, our findings suggest that LLMs can expand the share of decisions that managers base on established principles. An important caveat is that LLMs, unlike earlier knowledge-based systems, are much more reactive—they surface insights in response to what the user asks. Their effectiveness as a knowledge system therefore depends heavily on user expertise: knowing which questions to ask and how to connect concepts. Particularly for novice users, isolated prompts risk producing fragmented answers that mistake individual facts for genuine understanding.

Organizational implications. On the organization level, our findings imply that the cost of accessing foundational marketing knowledge within organizations drops substantially. This has two direct consequences. First, it levels the playing field: resource-constrained firms (including those with employees transitioning into marketing from other functions) can now access domain knowledge that previously required formal training or expensive consulting, so long as employees possess the ability to interact productively with these tools. Accordingly, firms may begin to place greater emphasis on cognitive traits that enable effective knowledge interaction rather than the a priori presence of domain knowledge. Second, it raises the return on existing marketing talent: experienced professionals become more effective when augmented with reliable on-demand access to foundational knowledge. A tempting but risky response is to treat LLM-accessible knowledge as "good enough" and reduce investment in training or talent.

Market-level implications. As foundational knowledge becomes universally accessible, the relative value of proprietary, context-specific marketplace knowledge, the kind that cannot be retrieved from a general-purpose LLM, is likely to increase. Firms that invest in developing unique customer insights, market intelligence, and organizational learning may find that these assets become more, not less, important. At the same time, if every firm has access to the same marketing principles, knowledge of these principles alone can no longer confer a competitive advantage. This may boost overall market performance but does not uniquely help any single firm, and may be a source of strategic homogenization (Krakowski et al., 2023; Wingate et al., 2025).

6.4 Implications for Marketing Educators

Finally, our findings carry direct and far-reaching implications for marketing education.

Assessment. Any assessment that a widely available LLM can pass with high marks no longer reliably differentiates student knowledge from AI-assisted output. Since our results show that LLMs achieve 83–94% accuracy across all 12 marketing topics, with near-perfect performance on recall and understanding, no area of foundational knowledge is exempt. Educators will need to shift toward assessments that target the integration of ambiguous information, judgment under uncertainty, and/or the defense of strategic positions in interactive settings.

Curriculum. Our results show that the skills traditionally tested in marketing education (recall, understanding, and increasingly even higher-order reasoning) are being conquered by LLMs. This raises a natural question: what should curricula focus on instead? Our findings point to two skill areas. First, *problem identification and framing*: our study evaluates LLMs on well-structured questions with clearly specified objectives and all relevant information provided. Real marketing problems are rarely presented this way. As LLMs become increasingly proficient at answering well-structured questions (a trajectory evident from GPT-3.5 to current reasoning models), the distinctive human contribution may shift toward deriving those questions from messy reality: identifying the underlying problem, deciding what information is needed, and framing the question appropriately. Second, *prompting and interaction*: even assuming perfect LLM knowledge, extracting value from it is a skill in its own right. The user must know which questions to ask, how to elicit connections across concepts, and how to integrate responses into a coherent understanding. These are skills that curricula will need to develop deliberately rather than assume they emerge through LLM use alone. Finally, marketing decisions are embedded in social contexts that require persuasion, negotiation, stakeholder management, and the ability to build trust, capabilities that remain fundamentally human.

6.5 Limitations

Our benchmark covers the established body of foundational marketing knowledge as codified in widely used academic textbooks. This knowledge has typically been developed, validated, and refined over extended periods before being incorporated into teaching materials. While this provides a well-established foundation for evaluation, it carries several limitations. First, it does not capture the tacit, experiential marketplace knowledge developed through real-world application (Wierenga, 2002), which is context-dependent and harder to evaluate at scale. Second, it does not speak to the more fluid frontier knowledge that is developed and disseminated through journal articles but has not yet been codified in textbooks. As Table 1 illustrates, rapidly evolving domains such as artificial intelligence are not yet well represented in our benchmark. Our evaluation therefore provides a snapshot of LLM capabilities against a largely established knowledge base.

Our evaluation also relies on closed-ended and short open-ended question formats. These formats offer clear advantages in scalability, which enabled coverage of the full breadth of the discipline, but they also have natural limitations. By design, our questions are restricted to relatively unambiguous cases in which, from the textbook authors' perspective, a single correct answer exists. This excludes situations involving genuine ambiguity, competing perspectives, or context-dependent trade-offs. Moreover, because our questions are isolated rather than interconnected, they can only partially capture a model's deeper, integrative understanding of how marketing concepts relate to one another. While our supplementary analysis of open-ended discussion questions speaks to this to some degree, more elaborate evaluation modes, such as structured interviews or longer open-ended case analyses, could assess capabilities that our formats cannot, though they are difficult to run at the required scale.

Finally, while we can quantify how accuracy varies across question characteristics, our analysis cannot always speak to the origin of these differences. For instance, the small residual topic-level differences we observe could reflect inherent difficulty, differential representation in training corpora, or compositional differences in question types. This interpretive limitation, however, does not affect our ability to control for these factors as potential confounders. Our main findings are estimated

conditional on topic and textbook fixed effects, which absorb any systematic differences in topic- or textbook-level difficulty, training data representation, or question composition.

6.6 Future Research

Our work opens several directions for future research. First, although we show that LLMs can provide reliable access to marketing knowledge, we do not examine how that knowledge is used in practice. Future research should investigate whether and when managers seek out such knowledge, how effectively they incorporate it into decision-making, and under what conditions doing so improves outcomes. The benefits of LLM support may also depend on characteristics of the user, such as prompting skill, critical thinking ability, or cognitive framing.

A related question concerns firms' strategic responses to this shift: whether LLMs enable new marketing knowledge investments that were previously too costly, complement existing investments in human expertise, or substitute for them. Determining when each of these responses enhances or erodes performance remains an open empirical question. A related avenue is to explore how organizations might encode and expose internal knowledge (such as the lived experiences of senior salespeople or brand managers) so that it can be leveraged through LLMs. It might be fruitful to revisit earlier attempts to develop knowledge-based systems (Rangaswamy et al., 1989; Burke et al., 1990), which might become revitalized through the capabilities provided by LLMs.

Second, as noted in our limitations, our benchmark captures a snapshot of established knowledge as codified in textbooks. Marketing knowledge, however, is not static. The discipline continually produces new findings, updates established frameworks, and occasionally overturns previously accepted insights. Whether LLMs correctly incorporate such updates, and equally important, whether they appropriately displace outdated knowledge, is an open question that our evaluation cannot address. Likewise, rapidly evolving topics such as artificial intelligence are underrepresented in current textbooks, and frontier knowledge disseminated through journal articles falls outside our benchmark entirely. Evaluating LLMs' marketing knowledge should therefore not be a one-time exercise but an ongoing effort.

Third, with the rise of AI agents that can autonomously plan, act, and interact with their environment, the next challenge extends beyond what LLMs know to how effectively they translate knowledge into action. There is early work on developing such benchmarks (Huang et al., 2025), which marketing research could build upon.

7 Conclusion

Marketing scholars have long worked on advancing the frontier of marketing knowledge, with debates around relevance and impact often focusing on whether we are studying the right problems in the right ways (e.g., Madan et al., 2023). A different question, which deserves equal attention, is: how can we ensure that the existing knowledge is available to the relevant stakeholders? In this light, large language models may represent a powerful tool for broadening our work's reach, and thereby a novel, and perhaps long-overdue, pathway for increasing the accessibility and practical impact of academic research.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the authors used ChatGPT and Claude in order to improve language and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

AMA (2025). Journal of marketing insights in the classroom. Accessed June 10, 2025, https://www.ama.org/journal-of-marketing-insights-in-the-classroom/.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.

Arora, N., Chakraborty, I., & Nishimura, Y. (2025). AI-human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators. *Journal of Marketing*, 89(2), 43–70.

Blanchard, S. J., Duani, N., Garvey, A. M., Netzer, O., & Oh, T. T. (2025). New tools, new rules: A

practical guide to effective and responsible GenAI use for surveys and experiments in research. *Journal of Marketing*, 89(6), 119–139.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for market research. SSRN preprint, submitted March 21. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4395751.

Brucks, M. & Toubia, O. (2025). Prompt architecture induces methodological artifacts in large language models. *PLoS One*, 20(4), e0319159.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2), 889–942.

Burke, R. R. (1991). Reasoning with empirical marketing knowledge. *International Journal of Research in Marketing*, 8(1), 75–90.

Burke, R. R., Rangaswamy, A., Wind, J., & Eliashberg, J. (1990). A knowledge-based system for advertising design. *Marketing Science*, 9(3), 212–229.

Burnap, A., Hauser, J. R., & Timoshenko, A. (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science*, 42(6), 1029–1056.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Ponde de Oliveira Pinto, H., Kaplan, J., Edwards, H., et al. (2021). Evaluating large language models trained on code. arXiv preprint, submitted July 7. https://arxiv.org/abs/2107.03374.

Chen, W.-L., Wei, Z., Zhu, X., Feng, S., & Meng, Y. (2025). Do LLM evaluators prefer themselves for a reason? arXiv preprint, submitted April 4. https://arxiv.org/abs/2504.03846.

Day, G. S. (1994). The capabilities of market-driven organizations. *Journal of Marketing*, 58(4), 37–52.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Organization Science*, 37(2), 403–423.

Deveau, R., Griffin, S. J., & Reis, S. (2023). AI-powered marketing and sales reach new heights with generative AI. McKinsey & Company, Los Angeles, CA.

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

Eisend, M. (2015). Have we progressed marketing knowledge? A meta-meta-analysis of effect sizes in marketing research. *Journal of Marketing*, 79(3), 23–40.

Garvey, A. & Blanchard, S. J. (2025). Generative AI as a research confederate: The LUCID methodological framework and toolkit for human-AI interactions research. SSRN preprint, submitted May 15. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5256150.

Germann, F., Lilien, G. L., & Rangaswamy, A. (2013). Performance implications of deploying marketing analytics. *International Journal of Research in Marketing*, 30(2), 114–128.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, 46(2), 107–119.

Glazer, R. (1991). Marketing in an information-intensive environment: Strategic implications of knowledge as an asset. *Journal of Marketing*, 55(4), 1–19.

Goli, A. & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*, 43(4), 709–722.

Grewal, D., Satornino, C. B., Davenport, T., & Guha, A. (2025). How generative AI is shaping the future of marketing. *Journal of the Academy of Marketing Science*, 53(3), 702–722.

Gui, G. & Toubia, O. (2023). The challenge of using LLMs to simulate human behavior: A causal inference perspective. SSRN preprint, submitted December 1. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4650172.

Hambrick, D. C. & Chen, M.-J. (2008). New academic fields as admittance-seeking social movements: The case of strategic management. *Academy of Management Review*, 33(1), 32–54.

Hartmann, J., Exner, Y., & Domdey, S. (2024). The power of generative marketing: Can generative AI create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1), 13–31.

Heitmann, M., Jansen, T., Reisenbichler, M., & Schweidel, D. A. (2025). EXPRESS: Picture perfect: Engaging customers with visual generative AI. *Journal of Marketing*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proc. 9th Int. Conf. Learning Representations*.

Hermann, E. & Puntoni, S. (2025). Empowering GenAI stakeholders. *Journal of the Academy of Marketing Science*, 53(3), 677–683.

Huang, K.-H., Prabhakar, A., Thorat, O., Agarwal, D., Choubey, P. K., Mao, Y., Savarese, S., Xiong, C., & Wu, C.-S. (2025). CRMArena-Pro: Holistic assessment of LLM agents across diverse business scenarios and interactions. arXiv preprint, submitted May 24. https://arxiv.org/abs/2505.18878.

Ipsos (2026). Marketing anchors: The case for capability in an era of transformation. Ipsos. https://www.ipsos.com/en-us/marketing-anchors-case-capability-era-transformation.

Joerling, M. (2026). Integrating GenAI interactions in marketing studies: A methodological guide. *International Journal of Research in Marketing*, 43(1), 28–47.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proc. 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611.

Jürgensmeier, L. & Skiera, B. (2024). Generative AI for scalable feedback to multimodal exercises. *International Journal of Research in Marketing*, 41(3), 468–488.

Kapoor, A. & Kumar, M. (2025). Frontiers: Generative AI and personalized video advertisements. *Marketing Science*, 44(4), 733–747.

Karpathy, A. (2024). Jagged intelligence. Twitter (July 25). https://x.com/karpathy/status/1816531576228053133.

Keon, J. W. & Bayer, J. (1986). An expert approach to sales promotion management. *Journal of Advertising Research*, 26(3), 19–26.

Kohli, A. K. & Haenlein, M. (2021a). Factors affecting the study of important marketing issues: Additional thoughts and clarifications. *International Journal of Research in Marketing*, 38(1), 29–31.

Kohli, A. K. & Haenlein, M. (2021b). Factors affecting the study of important marketing issues: Implications and recommendations. *International Journal of Research in Marketing*, 38(1), 1–11.

Korst, J., Puntoni, S., Purk, M., Smith, B., Colón, A., & Urbina-McCarthy, D. (2024). Growing up: Navigating GenAI's early years. AI at Wharton and GBK Collective, Philadelphia, PA.

Krakowski, S., Luger, J., & Raisch, S. (2023). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal*, 44(6), 1425–1452.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218.

Kumar, V. (2017). Integrating theory and practice in marketing. *Journal of Marketing*, 81(2), 1–7.

Kučera, H. & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Lehmann, D. R. (2014). Introduction to the special issue on theory and practice in marketing. *Journal of Marketing Research*, 51(6), 645–646.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The Winograd schema challenge. In Brewka, G., Eiter, T., & McIlraith, S., editors, *Proc. 13th Int. Conf. Principles of Knowledge Representation and Reasoning*, pages 552–561, Washington, DC. AAAI Press.

Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43(2), 254–266.

Lilien, G. L., Roberts, J. H., & Shankar, V. (2013). Effective marketing science applications: Insights from the ISMS-MSI practice prize finalist papers and projects. *Marketing Science*, 32(2), 229–245.

Ma, L. & Sun, B. (2020). Machine learning and AI in marketing—connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3), 481–504.

Madan, S., Johar, G. V., Berger, J., Chandon, P., Chandy, R., Hamilton, R., John, L. K., Labroo, A. A., Liu, P. J., & Lynch Jr, J. G. (2023). Reaching for rigor and relevance: Better marketing research for a better world. *Marketing Letters*, 34(1), 1–12.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint, submitted February 9. https://arxiv.org/abs/1802.03426.

Mohammadi, B. (2024). Explaining large language models decisions using Shapley values. SSRN preprint, submitted March 15. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4759713.

Moorman, C. & Miner, A. S. (1997). The impact of organizational memory on new product performance and creativity. *Journal of Marketing Research*, 34(1), 91–106.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

OpenAI (2025). Reasoning models – OpenAI API documentation. Accessed April 2026. https://developers.openai.com/api/docs/guides/latest-model.

Patwardhan, T. et al. (2025). GDPval: Evaluating AI model performance on real-world economically valuable tasks. arXiv preprint, submitted October 6. https://arxiv.org/abs/2510.04374.

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*, 40(2), 269–275.

Phan, L. et al. (2025). Humanity's last exam. *Nature*, 638, 132–138.

Rangaswamy, A., Eliashberg, J., Burke, R. R., & Wind, J. (1989). Developing marketing expert systems: An application to international negotiations. *Journal of Marketing*, 53(4), 24–39.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34), e2308950121.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2024). GPQA: A graduate-level google-proof Q&A benchmark. In *Proc. 1st Conf. Language Modeling*.

Reisenbichler, M., Reutterer, T., & Schweidel, D. A. (2026). Applying large language models to sponsored search advertising. *Marketing Science*, 45(1), 123–141.

Reisenbichler, M., Reutterer, T., Schweidel, D. A., & Dan, D. (2022). Frontiers: Supporting content marketing with natural language generation. *Marketing Science*, 41(3), 441–452.

Roberts, J. H., Kayande, U., & Stremersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing*, 31(2), 127–140.

Rossiter, J. R. (2001). What is marketing knowledge? Stage I: Forms of marketing knowledge. *Marketing Theory*, 1(1), 9–26.

Saxena, R., Gema, A. P., & Minervini, P. (2025). Lost in time: Clock and calendar understanding challenges in multimodal LLMs. arXiv preprint, submitted February 7. https://arxiv.org/abs/2502.05092.

Schauerte, N., Becker, M., Imschloss, M., Wichmann, J. R., & Reinartz, W. J. (2023). The managerial relevance of marketing science: Properties and genesis. *International Journal of Research in Marketing*, 40(4), 801–822.

Stremersch, S. (2021). The study of important marketing issues: Reflections. *International Journal of Research in Marketing*, 38(1), 12–17.

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025). Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6), 1446–1455.

Van Heerde, H. J., Moorman, C., Moreau, C. P., & Palmatier, R. W. (2021). Reality check: Infusing ecological value into academic marketing research. *Journal of Marketing*, 85(2), 1–13.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., et al. (2024). MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Proc. 38th Conf. Neural Information Processing Systems*, pages 95266–95290, San Diego, CA. Neural Information Processing Systems Foundation.

Wierenga, B. (1990). The first generation of marketing expert systems. Working paper, Rotterdam School of Management, Erasmus University.

Wierenga, B. (2002). On academic marketing knowledge and marketing knowledge that marketing managers use for decision-making. *Marketing Theory*, 2(4), 355–362.

Wierenga, B. (2021). The study of important marketing issues in an evolving field. *International Journal of Research in Marketing*, 38(1), 18–28.

Wierenga, B. & Van Bruggen, G. H. (1997). The integration of marketing problem-solving modes and marketing management support systems. *Journal of Marketing*, 61(3), 21–37.

Wingate, D., Burns, B. L., & Barney, J. B. (2025). Why AI will not provide sustainable competitive advantage. *MIT Sloan Management Review*, 66(4), 9–11.

Winter, S. G. (2009). Knowledge and competence as strategic assets. In Klein, D. A., editor, *The Strategic Management of Intellectual Capital*, pages 165–187. Routledge.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Appendix

A. Sample Questions from our Knowledge Base

Table A1 provides sample questions for each of our 12 marketing topics.

Table A1: Sample Questions for Each Marketing Topic (Part 1)

#	Topic Name	Question Text	Book Answer
1	Social Media Marketing	For a viral marketing campaign to work, individuals must be offered financial incentives to pass the message along to other consumers	false
2	Sales & Customer Value	A customer's profitability is judged on the basis of the lifetime stream of revenue and cost, not the profit from a particular transaction.	true
3	Product & Brand Management	Apple has churned out one cutting-edge product after another. It all started with the sleek, affordable Apple Macintosh, the first personal computer ever to feature a graphic user interface and mouse. That was followed by an Apple-led revolution in which groundbreaking Apple products such as the iPod, iTunes, the iPhone, and the iPad all created whole new categories where none previously existed. Apple is an example of _____. A) customer intimacy B) product differentiation C) operational excellence D) product leadership E) focus	d
4	Competitive Strategy	Which of the following is an advantage of nonprice competition? a. A firm can react quickly to competitive efforts. b. Market share becomes less important. c. A firm can build customer loyalty. d. Marketing efforts are completely eliminated. e. Pricing is no longer a factor.	c
5	Dynamic Market Analysis	Contingency plans should be based directly on _____. A) gap analysis B) benchmarking C) scenario analysis D) strategic planning	c
6	Market Research	There are three main types of research designs employed in marketing research: exploratory, descriptive, and conclusive.	false
7	Statistical Methods	In preparing categorical variables for analysis, it is usually best to _____. a. convert the categories to numeric representations b. convert the categories to binary, dummy variables c. combine as many categories as possible d. let them remain categorical	b

Table A1: Sample Questions for Each Marketing Topic (Part 2)

#	Topic Name	Question Text	Book Answer
8	Global/Local Marketing	A company should make several important decisions before deciding which markets to enter. These include all of the following EXCEPT _____. A) how many countries it wants to enter B) what volume of foreign sales it wants C) how many different products it wants to offer D) what its international marketing objectives are E) what its international marketing policies are	c
9	Communication Process	An ad for Maybelline age-minimizing makeup in Ladies' Home Journal magazine featured international film star Ziyi Zhang and offered readers a $1-off coupon to try the new makeup. In terms of the SMCR communication model, which of the following would be the best way for the source to measure feedback? A) the number of subscribers to Ladies' Home Journal B) the number of people who make up the target market C) the number of people who redeem the coupon D) the number of people who were exposed to the ad E) the number of people to whom Ziyi Zhang is an appealing spokesperson	c
10	Consumer Cognition	Neo-Freudian theorists believe that _____. A) consumption situations are extensions of the consumer's personality B) human drives are largely unconscious C) social relationships are fundamental to the formation and development of personality D) consumers are primarily unaware of their true reasons for making decisions E) consumer purchases are a reflection of an individual's personality	c
11	Marketing Models & Simulations	Correlating input variables in a simulation model accounts for which of the following? a. Input variables are often not independent in real-world situations. b. Output variables cannot be normally distributed unless inputs are correlated. c. Correlations ensure that all input variables have the same distribution. d. Correlations eliminate the need for multiple simulation iterations.	a
12	Marketplace Relationships	_____ occurs when a firm works with others inside and outside of the firm to bring more value to their customers. A) Targeting new customers B) Partner relationship management C) Customer brand advocacy D) Customer-engagement marketing E) Partnership marketing	b