Annals of "Dunarea de Jos" University of Galati
Fascicle I. Economics and Applied Informatics
Years XXXI – nº3/2025
ISSN-L 1584-0409 ISSN-Online 2344-441X
www.eia.feaa.ugal.ro

DOI https://doi.org/10.35219/eai15840409551

From SEO to AEO for Optimizing LLM Output in E-commerce Applications

Maria Cristina Enache★

ARTICLE INFO

Article history:
Received November 14, 2025
Accepted October 12, 2025
Available online December 2025
JEL Classification
D12, L86, D83

Keywords:
e-commerce, LLM, artificial intelligence

ABSTRACT

This paper proposes a methodological framework for transitioning from traditional Search Engine Optimization (SEO) to a new paradigm—Answer Engine Optimization (AEO)—driven by Large Language Models (LLMs) and multimodal Generative AI systems. The study contrasts the architectural philosophies of leading LLMs, such as GPT-4o and Claude, illustrating how Custom Instructions, Constitutional AI, and long-context design affect enterprise-grade deployment. It introduces a structured approach to product description generation and content optimization in e-commerce, emphasizing Semantic Clarity, Schema.org markup, and Extraction-Readiness as prerequisites for LLM citation and high-quality output. The framework ultimately positions LLM-optimized content as a structured, machine-interpretable data object, offering a foundation for next-generation e-commerce systems that integrate conversational commerce, multimodal diagnostics, and automated technical support.

1. Introduction

The landscape of online discovery is undergoing a significant transformation. For years, Search Engine Optimization (SEO) has been the cornerstone of digital marketing, focusing on optimizing websites and content to achieve prominent rankings in search engine results, primarily Google. This intricate dance between content creators and search algorithms has spawned a vast industry. However, the advent of generative AI, large language models (LLMs), and sophisticated AI chatbots is rapidly reshaping this paradigm.

The rapid evolution of Generative AI (Gen AI) has fundamentally redefined the operational and strategic landscape of digital commerce. As Large Language Models (LLMs) transition from experimental tools to core infrastructural components, organizations increasingly rely on them to accelerate content production, automate technical support, and enhance user interaction across omnichannel systems. This shift has exposed a structural limitation in traditional Search Engine Optimization (SEO), whose methods were originally designed for crawler-based indexing rather than AI-mediated interpretation. In contemporary Search Generative Experiences (SGE), information is not merely retrieved but synthesized, necessitating a paradigm in which content must be architected for machine understanding rather than human scanning alone. At the center of this transition lies a deeper need to understand the architectural foundations of modern LLMs. Models such as GPT-4o and Claude embody contrasting design philosophies—generalized versatility versus constitutional alignment—that directly influence their behavior, reliability, and applicability in enterprise contexts. Their capabilities in long-context reasoning, multimodal processing, and persistent instruction-following enable entirely new forms of content optimization, interaction design, and automated decision-making. These advancements catalyze the emergence of Answer Engine Optimization (AEO), a methodology that reframes content as a structured, machine-interpretable data object. AEO prioritizes semantic clarity, explicit markup, disambiguation, and extraction-readiness—allowing LLMs to cite, compare, and synthesize information with high fidelity. In e-commerce, where product attributes, warranties, and technical specifications must be both accurate and contextually consistent, this shift represents not just an optimization technique but an infrastructural requirement. Simultaneously, enterprises face increasing pressure to deploy these systems efficiently.

Multi-Model architectures—where several LLMs are orchestrated through a centralized routing layer—offer a pragmatic solution by balancing cost, latency, and task-specific performance. Coupled with Retrieval-Augmented Generation (RAG) pipelines, these architectures unlock high-confidence reasoning over extensive product documentation, manuals, policies, and multimedia assets. In this context, the integration of Gen AI into e-commerce is no longer constrained to copywriting or search visibility. It extends across the entire

★Dunarea de Jos University of Galati, Romania. E-mail address: mpodoleanu@ugal.ro (M. C. Enache).

product lifecycle: from pre-sales comparison assistance to post-sales troubleshooting, multimodal diagnostics, and automated L3 technical support.

This paper proposes a comprehensive framework for implementing SEO-to-AEO transformation, describing the architectural, linguistic, and operational competencies required to optimize LLM output at scale. Through comparative model analysis, system design patterns, and an applied case study, the framework aims to provide a foundational methodology for organizations seeking to adapt their content ecosystems to the AI-first search paradigm.

The successful utilization of Large Language Models (LLMs) in commercial and enterprise applications is critically conditioned by their fundamental design philosophy and underlying model architecture. The ChatGPT model, underpinned by the GPT-4o Transformer architecture, prioritizes generalized versatility and optimization for Conversational AI interactions. Its robust API platform facilitates cross-platform integration and scaling of Natural Language Generation/Understanding (NLG/NLU) functionalities.

A key technical advantage lies in the implementation of Custom GPTs and Custom Instructions, serving as a mechanism for Persistent Prompt Engineering (PPE). This allows users to inject pre-defined contextual and stylistic metadata, ensuring the model's behavioral alignment with specific enterprise requirements. Furthermore, its real-time web access capability allows for the integration of dynamic and up-to-the-minute data, mitigating the limitations of its static Knowledge Cutoff.

In contrast, Claude, developed by Anthropic, is fundamentally grounded in Constitutional AI, a training methodology that enforces a set of ethical principles. This architectural foundation actively limits open behavioral drift, making the model ideal for regulated environments where strict ethical compliance is non-negotiable. Technically, Claude excels in document analysis and extensive summarization due to its remarkably large context window size. This technical attribute is essential for maintaining referential coherence and long-term memory throughout extended dialogues or during the processing of vast textual corpora, preventing the common Lost-in-the-Middle Phenomenon.

2. Integrating Gen AI in Product Description Generation (Conversational Optimization)

The evolution from traditional SEO mandates a shift to Generative Engine Optimization (AEO), which focuses on optimizing content specifically to be easily cited and synthesized by LLMs. This requires prioritizing Semantic Clarity and, crucially, implementing Semantic Markup (Schema.org). This practice involves embedding structured data, such as JSON-LD for FAQSchema, directly within the product pages. For instance, to ensure citation, companies embeds explicit data structures for warranty inquiries in the home appliance sector, such as the following, to maximize LLM confidence in comparison and citation:

```
{
 "@context": "https://schema.org",
 "@type": "FAQPage",
 "mainEntity": [{
  "@type": "Question",
  "name": "What is the standard warranty period for this model?",
  "acceptedAnswer": {
   "@type": "Answer",
   "text": "The standard warranty period for this appliance model is 24 months, subject to authorized installation."
  }
 }]
}
```

This structural adherence ensures that LLMs can perform high-fidelity information extraction. Furthermore, Conversational GEO SEO involves the strategic insertion of localized signals to maximize visibility in AI responses to localized intent queries. Despite these technical enhancements, human validation remains indispensable due to the persistent risk of factual hallucinations and the necessity of injecting proof points of expertise (E-E-A-T).

The deployment of advanced conversational assistants and chatbots is currently migrating towards multi-AI architectures, a necessity driven by the mandate to balance optimal performance (speed and precision) with operational cost efficiency. These architectures do not rely on a single language model but aggregate the APIs of several competing LLMs to optimize model selection based on the specific requirements of the task. The central component of this architecture is the Router Layer. It operates as a request arbitration system, evaluating each user query based on established criteria before allocating it to the appropriate LLM.

The routing decision is made using a matrix that typically includes:

- Task Complexity: Queries requiring extended contextual coherence or deep text analysis (such as cross-referencing sections within lengthy manuals) are directed to models with a superior context window, like Claude 3.5 Sonnet. Conversely, tasks demanding creativity, low latency response speed, or code generation are routed to more agile models, such as GPT-4o.

- Cost Efficiency: For basic, repetitive tasks (L1 Support or short greeting messages), the Router selects models with a lower marginal token cost to maintain a sustainable Cost of Goods Sold (COGS).
- Model Specialization: Models that excel in certain domains (e.g., a finetuned model for specific appliance technical language) are given priority for those types of queries.

This Router Layer essentially uses an initial prompt sent to a faster base model (or a smaller, specialized classification model) to determine the user's intent and the structural complexity of the request before making the final call to the most expensive or most powerful model.

When a query necessitates access to private documents (manuals, policies, technical sheets), the system enters RAG mode. Within the multi-model architecture, the Router Layer also dictates which LLM will execute the generation phase of the RAG process:

- RAG for Technical Data Synthesis: If the query is a simple specification lookup (e.g., "What is the noise level of this refrigerator?"), the fast model is selected to extract and synthesize information from a single data chunk.
- RAG for Multi-Document Coherence (High-Coherence RAG): If the query is complex, requiring the simultaneous analysis of multiple indexed documents (e.g., correlating warranty terms with installation instructions), the Router chooses the model with the largest Context Window (such as Claude). This decision prioritizes inference quality and minimizes the risk of error, even if it entails a higher processing cost.

In essence, the multi-model architecture transforms the conversational system from a monolith into an AI Orchestrator, capable of applying the principle of "the right model for the right job" at an enterprise scale. A conceptual example of the Python routing logic, based on a derived complexity score (though real-world logic is more nuanced), illustrates the mechanism:

```
def route_query(user_query, complexity_score, document_type):
  if document_type == "TechnicalManual" and complexity_score > 0.7:
    # Utilizes the Long Context Window model (e.g., Claude) for RAG
    return "Claude_Opus_for_Deep_Analysis"
  elif "creative" in user_query.lower() or complexity_score < 0.3:
    # Utilizes the fast, versatile model (e.g., GPT-4o) for speed and cost
    return "GPT_4o_for_Low_Latency"
  else:
    # Default routing or routing to a low-cost model
    return "Low_Cost_Model"
```

This architectural stratification is essential for e-commerce companies like a digital retailer that deliver the necessary accuracy for complex L3 technical support while efficiently managing costs for routine pre-sales inquiries.

Personalization with Gen AI extends beyond mere content generation to the stylometric adaptation of the output to the user's pre-existing narrative voice (brand voice). The LLM's value resides in its ability to function as a sophisticated stylistic editor, amplifying the existing voice rather than creating one ex nihilo.
The technical foundation for this process is stylometric analysis, where the user's writing style is decomposed into quantifiable parameters that the model can process and replicate. These parameters include: Lexical Density (the ratio of content words to function words), Syntactic Complexity (measured by sentence length, clause frequency, and the Flesch–Kincaid index), Tonal Semantics (identifying subjective language associated with voices like formal, ironic, or emphatic), and Rhythmic Variation (the measured alternation between short and long propositional structures).

Stylistic Imitation is typically achieved without costly model re-training (finetuning). Instead, the model is provided with reference texts (few-shot examples) written by the user, embedded within Long Prompts. The instruction within the prompt directs the LLM to perform a comparative stylistic analysis on the input examples, specifically isolating the user's characteristic lexical tics or favored syntactical patterns. Subsequently, the model is commanded to replicate the identified style and structure in the generation of new content. For an e-commerce brand this ensures that all communication—from technical support answers to social media posts—maintains a unified brand voice, adhering, for example, to a consistently authoritative yet approachable tone.

At the interface level, this personalization is rigidly enforced through the utilization of Custom Instructions (or System Messages in API implementations). These function as persistent stylistic constraints, defining the model's persona, context, and, crucially, specific output preferences across all subsequent interactions. These instructions override the general model behavior, defining the LLM's role—for instance: "You are the Lead Technical Copywriter for the retailer. Maintain a tone that is precise and authoritative. Never use first-person pronouns or exclamation points, and restrict the use of passive voice to under 10% of the total

generated text." This mechanism ensures that the desired style remains enforced without needing to restate the preferences in every user prompt.

3. Implementing Generative Engine Optimization (GEO/AEO) in E-commerce

	SEO Search Engine Optimization	AEO Answer Engine Optimization	GEO Generative Engine Optimization
Goal	Rank higher on traditional search engine results page	Secure featured snippets and voice search answers	Influence Ai-generated summaries and answer boxes
Primary focus	Keywords, backlinks, technical SEO	Structured answers, schema markup, FAQ	Natural language, semantic richness, context-aware content
Optimized For	Traditional engine like Google	Featured snippets, voice search and direct answers	Ai platforms like Google SGE, ChatGPT
Content Style	Informative, long-form, keyword-optimized	Concise, structured, FAQ-base	Conversational, fact-rich, conceptually connected

Figure 1. Conceptual Transition from Classical SEO to Answer Engine Optimization (AEO)

In recent years, many e-commerce platforms have faced a persistent challenge of low visibility within the Search Generative Experience (SGE). The strategic shift involved transforming content into a structured and comparable data object. This included implementing stratified JSON-LD markup to highlight quantifiable technical properties (e.g., Energy Class, Acoustic Emission) and refactoring content into an Extraction-Readiness Q&A format, where the core technical answer is isolated into a primary, autonomous sentence. E-E-A-T was established by attributing technical responses to an Author Tag with verifiable qualifications (e.g., Certified Service Technician), while Conversational GEO SEO optimized for localized services (e.g., "authorized installation for washing machines in the Bucharest metropolitan area"). This strategy successfully increased Citation Authority and converted low-quality traffic into high-intent users, evidenced by a significant reduction in the bounce rate.

A Full GEO/AEO strategy utilizes the specialized capabilities of LLMs for visual assets and complex technical support, encompassing the entire product lifecycle.

Stage (Phase)	Primary AEO Objective	Key Actions
Intent Research	Identifying the exact questions and the expected answer format required by AI.	Mapping out questions (not just keywords), defining the necessary Direct Answer.
Content Structuring	Ensuring a format that is easily extractable and understandable by AI models.	Front-loading the concise answer at the beginning of the section, heavy use of bulleted/numbered lists, and question-based headings.
Technical Optimization	"Translating" the content structure for search bots using metadata.	Implementing Schema Markup (e.g., FAQPage, HowTo) to tag answer types and author data (E-E-A-T).
Authority and Distribution	Increasing the trust and authority of the content to be chosen as a citation source.	Gaining brand mentions and citations from reliable sources; ensuring consistency of essential information across all channels.
Measurement and Iteration	Monitoring performance within the AI environment and continuously adjusting the strategy.	Tracking AI Visibility (how often the content is cited in AI answers) and regularly updating factual data.

This table summarizes the main stages of the Content Pipeline for Answer Engine Optimization (AEO), focusing on making content optimal for citation by AI and answer systems (like AI Overviews).

Multimodal models require an advanced approach to visual optimization beyond simple image tagging. An online commerce platform that manages an extensive portfolio of technical products uses the Gemini API for image analysis to facilitate post-sale assistance, a crucial differentiator in the appliance sector. For instance, the vision processing capability allows a customer to upload a photograph of a broken component (e.g., a specific filter or hinge) or a model identification plate. The model processes the image, cross-references it with the company's indexed visual database, and returns the exact Stock Keeping Unit (SKU), effectively transforming the visual input into a direct transactional data vector for spare parts ordering. Furthermore, the video content analysis (VCA) capability is used to generate Structured Video Data, where demonstrations of noise levels or energy-saving features are tagged to specific timestamps and associated with their technical metrics (e.g., 42dB), ensuring AEO validation through visual media.

Component	Sub-Elements	Role / Purpose	Why It Matters in E-commerce AI
1. User / Client Layer	Web interface, Mobile app, Chat widget, API Gateway	Receives user queries and sends them to Router Layer	Ensures consistent omnichannel AI behavior
2. Router Layer (Core Orchestration Engine)	Intent & Complexity Classifier Routing Rules Engine Cost/Latency Monitor	Analyzes query type, complexity, and cost constraints to decide which model should answer	Enables "right model for the right job" → reduces cost, increases accuracy
3. Model Pool (LLM Layer)	GPT-4o (creative, fast) Claude 3.5 (long-context reasoning) Low-cost model (L1 support) Domain-specialized model	Executes the user query using the model best suited for its complexity	Ensures performance efficiency across different query types
4. RAG Subsystem (when documentation is needed)	Vector Database Retriever Context Builder	Retrieves relevant passages from manuals, policies, and product docs and builds grounded context	Enables high-confidence technical support and policy reasoning
5. Post-Processing Layer	Output Validation (guardrails) Style/Brand Consistency Formatting JSON/E-mail/Chat response formatting	Refines and validates LLM output for accuracy, tone, and safety	Prevents hallucinations and ensures brand-compliant communication
6. Analytics & Feedback Loop	Query Logs Cost Tracking AEO Visibility Metrics	Monitors performance, accuracy, and model usage patterns	Supports continuous optimization and regulatory compliance

Figure 2. Core Components of a Multi-Model Routing Layer Architecture

For mission-critical support involving voluminous technical documentation (100+ page manuals, warranty booklets, and installation guides), a digital retailer leverages the superior Long Context Window of Claude Opus/3.5 Sonnet within a Retrieval-Augmented Generation (RAG) architecture.

This allows the system to execute complex cross-referencing queries that require synthesizing information from multiple, separate source documents to determine liability and service eligibility. This capacity for multi-document coherence is vital in the regulated appliance sector. A complex query executed by the RAG system to derive a synthesized policy answer would be structured similarly to: QUERY: "Synthesize the rules from the 'Warranty Booklet' (Document A) and the 'Installation Guide' (Document B). Specifically, if the customer's washing machine was installed without an approved anti-vibration mat (as required by Document B), does this installation non-compliance nullify the extended 5-year parts warranty offered in Document A?"

This level of precision and referential coherence automates L2/L3 technical support, allowing the system to deliver instant, high-confidence answers that correlate technical specifications with policy compliance, significantly boosting customer satisfaction and operational efficiency.

4. Conclusion

Transitioning from SEO to AEO marks a structural shift in how digital content is interpreted and surfaced. The study demonstrates that visibility in AI-mediated search experiences depends not merely on

keyword relevance but on a content, architecture optimized for interpretability by LLMs—structured, unambiguous, and validated through machine-readable markup.

Multi-Model architectures provide a scalable and cost-efficient foundation for enterprise conversational systems. By allocating queries to the right model based on complexity, context window, and cost, organizations can achieve high-precision reasoning without incurring unnecessary operational expenses. RAG pipelines further enhance reliability by grounding responses in proprietary documentation.

Persistent Prompt Engineering and stylometric analysis enable consistent brand voice across all generated outputs. Rather than functioning as ex nihilo generators, LLMs become stylistic editors, capable of replicating nuanced narrative signatures when supplied with reference texts and well-designed custom instructions.

GEO/AEO strategies extend beyond content creation into the full product lifecycle. The case study shows that structured data, multimodal interpretation (e.g., image-based SKU identification), and long-context policy reasoning allow e-commerce platforms to integrate pre-sales, post-sales, and technical support into a unified AI orchestration layer.

Despite automation, human oversight remains essential. Factual hallucinations, ambiguous specifications, and legal implications necessitate continuous validation, especially in regulated sectors such as home appliances and consumer safety.

The proposed framework positions LLM-aligned content as a strategic asset. Organizations that adopt AEO-aligned architectures early will gain an advantage in SGE ecosystems, improving both customer experience and operational efficiency while laying the groundwork for fully autonomous conversational commerce systems.

References

1. *Haidar, Illia. (2024). Applications of Artificial Intelligence in E-Commerce. Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023. 5. 32-38. 10.60087/jaigs.v5i1.151.*
2. *M. I. Ahmed, Understanding the artificial intelligence implementation for allocating an order to a seller among multiple sellers who sell the same product, in: K. Kang, F. Namisango (Eds.), E-Service Digital Innovation, IntechOpen, Rijeka, 2022*
3. *J. V. Louis, N. Noerlina, D. H. Syahchari, Digital business transformation: Analysis of the effect artificial intelligence in e-commerce's product recommendation, Advanced Information Systems 8 (2024) 64–69*
4. *K. Kang, X. Wang, W. Yang, The analysis of social e-commerce with artificial intelligence, Applied and Computational Engineering 47 (2024)*
5. *Andrea Polonioli, "Conversational Commerce: The Future of E-Commerce Customer Experience," Coveo, 2024. [Online]. Available: https://www.coveo.com/blog/conversational-commerce/*
6. *Cotinga, "E-Commerce Analytics for Customer Acquisition and Retention: A Data-Driven Approach," Cotinga Consulting, 2024 https://www.cotinga.io/blog/ecommerce-analytics-for-customer-acquisition/*