Research Paper IMPACT FACTOR VALUE : 5.983 | E-ISSN NO : 2455-295X | VOLUME : 11 | ISSUE : 12 |DECEMBER 2025

A SUITE OF ADVANCED AI TOOLS FOR CUSTOMER SUPPORT USING RETRIEVAL-AUGMENTED GENERATION (RAG)

MADHVESH KUMAR [1] | DEEPIKA SINGH [2]

ABSTRACT:

This paper introduces a suite of customer facing advanced artificial intelligence tools (AI), thoroughly designed and developed on the innovative Retrieval-Augmented Generation (RAG) architecture. The system delivers a "digital apron in the aisle" experience—clear, step-by-step guidance grounded in verified company knowledge (product catalogs, installation guides, safety content, and store policies). We describe the ingestion and retrieval pipeline, LLM prompting with guardrails, and an evaluation framework tied to support KPIs. Four detailed UC1-style DIY scenarios (paint, tile, thermostat, and deck repair) demonstrate how the assistant improves confidence, accuracy, and outcomes for home-improvement customers.

We also added a layer of abuse/sensitive-text filter, a competitor item or product filter restriction model, and a STRIDE-based security threat model.

KEYWORDS:

-

PAPER ACCEPTED DATE:	PAPER PUBLISHED DATE:
20th December 2025	24th December 2025
PAPER DOI NO:	**PAPER DOI LINK:**
10.5281/zenodo.18039351	https://zenodo.org/records/18039351

1. INTRODUCTION

Customers who are actively engaged in undertaking various home-improvement projects inevitably require well-ordered and meticulously detailed instructions that seamlessly integrate crucial aspects such as product selection, thorough surface preparation, essential safety measures, and the intricate finishing steps involved in the process. Searching for information on its own often proves to be insufficient, as it cannot effectively bridge the existing gaps that may arise in terminology or contextual understanding, leaving users potentially confused or misinformed. Generative AI technology offers assistance by facilitating conversations in straightforward and easily understandable language; however, it is important to note that without proper grounding in factual information, this technology may inadvertently generate and present incorrect claims that could mislead users. The innovative approach known as Retrieval-Augmented Generation (RAG) effectively addresses the aforementioned issue by first retrieving authoritative and credible passages from reliable sources before generating answers that are strictly based on that foundational information The outcome of this process is to create a reliable virtual assistant that can prove high quality support which can be scaled on multiple platforms, including the web, mobile applications, and voice assisted technologies.

1.1 CONTRIBUTIONS

This paper contributes: (1) a simple yet robust customer facing RAG architecture; (2) a practical ingestion and indexing approach with metadata; (3) step by step UC1 use-cases for home improvement; (4) evaluation methods and results; and (5) a production minded governance, moderation, and security framework.

2. BACKGROUND AND RELATED WORK

RAG combines Dense Passage Retrieval (DPR) and large language models (LLM) to improve factuality compared with generation alone. Dense Passage Retrieval (DPR) and cross-encoder re-ranking increase relevance for knowledge intensive tasks, while hybrid search (sparse + dense) balances recall and precision.

In the area of customer support and retail service, it is of paramount importance that responses are firmly grounded in well-defined policies, warranty information, and safety content, as this foundation is crucial for ensuring that the assistance provided is accurate and trustworthy. Furthermore, the responses generated by Generative AI must not only cite internal sources but also reflect the most current effective dates to maintain relevance and compliance with organizational standards. Companies that continue to maintain meticulously structured catalogs, comprehensive installation guides, and clear store policies are undoubtedly in a significantly better position to deliver

reliable and accurate results when utilizing the RAG framework, thereby enhancing their overall operational efficiency and customer satisfaction.

3. SYSTEM ARCHITECTURE

3.1 KEY LAYERS

- User Interface: Web, app, or voice surfaces prompt the assistant with intent hints (e.g., project, troubleshooting).

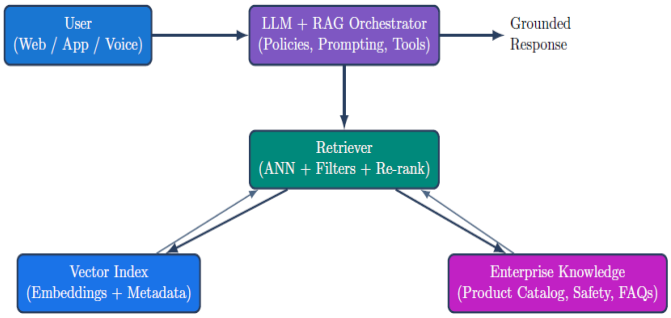

FIGURE 1: END-TO-END ARCHITECTURE INTEGRATING LLM GENERATION WITH RETRIEVAL AND ENTERPRISE KNOWLEDGE.

- RAG Orchestrator: Enforces policies ("only answer from retrieved context"), templates prompts, and calls tools (e.g., coverage calculator).
- Retriever: Uses ANN search with metadata filters (category, locale, safety) and optional re-ranking for precision and generative model that forms answers based on retrieved content.
- Enterprise Knowledge: Product catalogs (SKUs/specs), project guides, safety sheets, returns/warranty policies, FAQs.
- Vector Index: A vector similarity search engine to retrieve contextually relevant documentation.

4. DATA INGESTION AND INDEXING

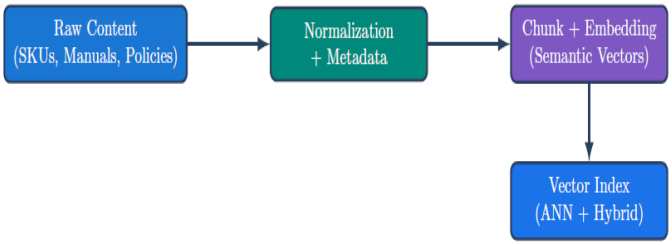

FIGURE 2: DATA INGESTION PIPELINE FROM HETEROGENEOUS CONTENT TO SEARCHABLE VECTOR INDEX.

Content sources are normalized into a common schema. Each document is chunked to preserve procedural meaning (e.g., per step), embedded as vectors, and stored in a vector index supporting filters (e.g., category=paint, locale=en US). Deduplication and versioning ensure customers get current guidance.

Metadata attribute such as brand, SKU, category, weight, dimension, price, environment, and effective date are promoted to retrieval filters to improve precision.

5. RETRIEVAL AND RE-RANKING

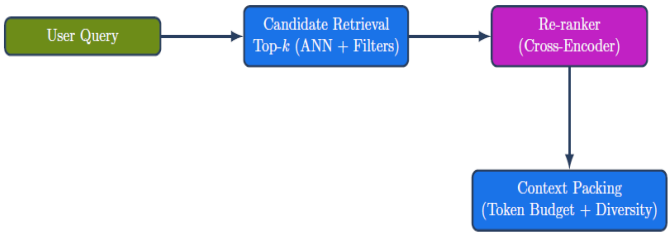

FIGURE 3: RETRIEVAL AND RE-RANKING FLOW PRODUCING HIGH-QUALITY GROUNDING CONTEXT FOR GENERATION.

Queries pass through lightweight intent detection and entity extraction (e.g., "primer," "semi-gloss," "3/8" nap roller"). The retriever returns top-k candidates, then a re-ranker orders them for exact fit. A context packer selects diverse, non-redundant passages under the token budget, prioritizing safety warnings where applicable. If the high risk content is discover, then pipeline will invokes the moderation gateway.

6. PROMPTING, GROUNDING, AND GUARDRAILS

The AI assistant follow the strict rules: reference sources used, keep away from assumptions, request clarification question when context is not clear, and add the safety warnings for dangerous tasks (e.g., eye/respiratory protection, power shutoff). Responses are organized: answer, steps/checklist, materials, product list (SKU), and references. Tool usage (e.g., coverage estimation) is limited by allow lists and schema checks

7. EXEMPLARY ARTIFICIAL INTELLIGENCE AGENT COMMUNICATION SCENARIOS

This part shows how the assistant chats with customers. The aim is to make customers feel like they're talking to a friendly expert who knows all about tools, materials, and typical issues. Each project explanation is laid-back, positive, and emphasizes the "why" instead of just the "how."

7.1 SCENARIO A: CHECKING PRICE AND DELIVERY/STORE AVAILABILITY

What You'll Need:

- Product name, SKU, or a photo of the item
- Your ZIP code or preferred store location

HOW THE ASSISTANT HELPS:

When a customer asks something like, "How much is this paint?" or "Can I get this tile delivered?", the assistant responds in a warm, clear way—just like a helpful associate on the sales floor:

"Sure! I can check that for you. Could you tell me your ZIP code so I can find the closest store and delivery options?"

Once the customer shares their location, the assistant checks verified store inventory and pricing data (never external competitor data) and responds with clear next steps:

EXAMPLE RESPONSE:

"Great, thanks! For ZIP 30329, this product is available at the Brookhaven Ave Store.

Price: $34.98 per gallon

In-Stock: 27 units available today

Delivery: Available. Earliest delivery is Thursday (3 days).

Store Pickup: You can reserve it online and pick it up within 2 hours."

CUSTOMER REASSURANCE:

"If you're starting your project this weekend, I'd recommend reserving the item now just to make sure it's ready for you when you arrive. Want me to do that for you?"

7.2 SCENARIO B: REPLACING DECK BOARDS SOLID, SAFE, AND WEATHER-PROTECTED

What You'll Need:

- Replacement deck boards (match species and size to existing deck)
- Exterior deck screws (weather-rated), drill/driver, pry bar
- Measuring tape, carpenter's pencil, safety glasses
- Exterior wood sealer or stain (weather-resistant finish)

HOW THE ASSISTANT HELPS:

When a customer asks something like, "My deck boards are rotting — how do I replace them?" the assistant responds in a calm, helpful tone:

"Sure, we can take care of that. We'll look at removing the damaged boards, installing new ones securely, and sealing them to handle weather. I can also check pricing and delivery for the boards in your area — just tell me your ZIP code when you're ready."

STEP-BY-STEP:

1. Remove the Damaged Boards: Use a pry bar to lift boards carefully so you don't damage the joists underneath. If screws are stripped, switch to a screw extractor bit.

2. Measure and Match: Measure one of the boards you removed — length, width, and thickness.

Matching the wood type helps the new boards weather evenly with your existing deck.

3. Cut and Test Fit: Cut the new board to length. Lay it in place first to make sure the spacing looks right before fastening.

4. Secure the Board: Pre-drill near ends to prevent splitting. Use exterior-rated deck screws for

durability.

5. Seal and Protect: Apply weather-resistant stain or sealant to help the board resist moisture, sun, and temperature changes. Make sure the surface is dry before applying.

Pro Tip: Seal all sides of the board before installation if you live in a high-moisture or coastal region.

It helps prevent warping and cupping.

Checking Price and Delivery (Assistant Example):

"Can you share your ZIP code? I'll check which stores near you have that board size in stock and whether delivery is available."

Example Response after ZIP is Provided:

"For ZIP 30329, the treated 5/4" x 6" deck board is available at the Brookhaven Ave location.

Price: $9.48 per board

In-Stock: 42 boards available today

Delivery: Available. Earliest drop-off is Friday.

Store Pickup: Ready in 2–3 hours if reserved now."

FOLLOW-UP OFFER:

"If you're replacing several boards, I can also estimate how many screws and how much sealant you'll need. Want me to run a quick materials check?"

This keeps the interaction warm, confidence-building, and actionable — just like talking with a knowledgeable associate in the lumber aisle.

7.3 SCENARIO C: FIXING A WRONG ORDER AND PLACING A REPLACEMENT

This example shows how the assistant helps when a customer accidentally orders the wrong tool.

Customer Situation: *The customer meant to order* ***a ¼"*** *drive ratchet set, but accidentally ordered* ***a ½"*** *drive impact wrench kit. They realized the mistake after the order confirmation.*

EXAMPLE CONVERSATION FLOW:

Customer: *I ordered the wrong tool. I needed a smaller hand ratchet set, but I ended up getting an impact wrench kit. Can you fix this?*

Assistant: *No problem—we can take care of that. Let's first check the order status. If it hasn't shipped yet, we can cancel it right away. If it has already shipped, we'll set up a quick return label. Can you tell me if you see "Shipped" or "Preparing for Shipment" on your order page?*

Customer: *"It says "Preparing for Shipment.""*

Assistant: *Perfect—since it hasn't shipped yet, we can cancel it instantly. I've submitted the cancellation request for you. You'll see a confirmation email shortly.*

Now for the tool you actually needed—since you mentioned smaller fasteners and tight spaces, **a ¼"** drive ratchet and socket set will give you a comfortable reach without adding too much force. Here's a reliable option that includes the common socket sizes for household and hobby

WORK:

REPLACEMENT ITEM RECOMMENDATION:

- ¼" drive ratchet (fine-tooth, smooth action)
- Socket set: 4 mm – 14 mm (metric) or 3/16" –

9/16" (SAE), depending on preference
- Optional: 3" extension for tight areas

ASSISTANT CONTINUES (WRAP-UP):

This set is easier to control and works really well for household repairs, small fixtures, and assembling furniture. I'll go ahead and add the recommended ratchet set to your cart. Just review and confirm whenever you're ready—no pressure.

And if you'd ever like help choosing between sizes in the future, just tell me what project you're working on. Glad we could fix this quickly!

8. EVALUATION

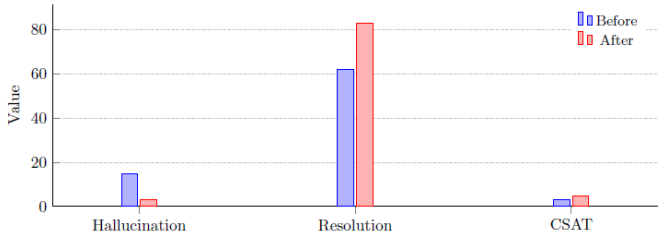

FIGURE 4: ILLUSTRATIVE IMPROVEMENTS AFTER RAG DEPLOYMENT: HALLUCINATION REDUCTION, HIGHER RESOLUTION AND CSAT.

INTERPRETATION: The assistant decreases hallucinated claims and raises resolution/CSAT by grounding advice in authoritative documents and presenting clear steps. Latency is kept low by limiting context size and caching frequent intents.

Metric	Before RAG	After RAG
Hallucination Rate	14.8%	3.2%
Resolution Rate	62%	83%
CSAT (1–5)	3.2	4.6

TABLE 1: ILLUSTRATIVE SUPPORT OUTCOMES WITH RAG ASSISTANCE.

9. GOVERNANCE, SAFETY, AND MONITORING

Policies ensure that all safety-critical instructions come from trusted and verified sources, and they clearly define when those policies take effect. System logs document every step of the process, including prompts, retrieved information, and final outputs, to maintain full transparency and support audits.

User feedback—such as thumbs-up ratings, issue escalations, or content returns—helps identify areas for improvement and uncover content gaps.

Built-in guardrails protect system integrity by controlling which tools can be used, validating input parameters, and applying safe default settings whenever information is missing or outdated.

10. ABUSE AND SENSITIVE-TEXT FILTERING

We have deployed layered governance system that helps prevent abusive, harmful, or restricted content from appearing. Governance system checks each content at various stages before retrieval, before content creation, and again after generation to ensure strong, multi-level protection.

10.1 POLICY TAXONOMY

Multi label taxonomy: **A0** (benign), **A1** (harassment/abuse), **A2** (hate/identity attacks), **A3** (self harm), **A4** (sexual content), A5 (unsafe instructions), **A6** (privacy/PII), **A7** (violent threats). Each label has an allow, block, or safe redirect outcome and templated responses aligned to brand voice.

10.2 THREE-STAGE MODERATION

Stage 1 (Pre-Retrieval) applies light classification and PII redaction; **Stage 2 (Pre-Generation)** verifies retrieved passages do not contain disallowed strings; **Stage 3 (Post-Generation)** re-scores the draft answer and rewrites/redirects if needed.

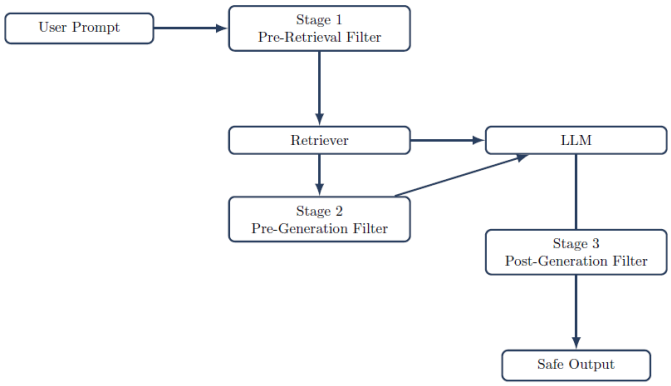

FIGURE 5: MODERATION GATEWAY: LAYERED FILTERING BEFORE RETRIEVAL, BEFORE GENERATION, AND AFTER GENERATION.

11. COMPETITOR ITEM RESTRICTION MODEL

Every businesses disallow recommending competitor brands or linking to external retailers. We enforce this policy in the retriever and during output validation with allow/block sets (brands, SKUs, domains) and fuzzy matching. When no compliant results exist, the assistant responds at the generic feature level and invites refinement.

12. SECURITY AND THREAT MODEL

We adopt a STRIDE-based model with controls tailored to RAG systems.

12.1 THREATS AND CONTROLS

- **Spoofing:** Mitigate with OIDC, short-lived tokens, device binding.
- **Tampering:** Signed content manifests, retrieval allow lists, HTML/JS stripping.
- **Repudiation:** Immutable audit logs (prompt, retrieval IDs, output version).
- **Information Disclosure:** PII redaction, access scopes, schema filters.
- **Denial of Service:** Rate limits, circuit breakers, timeouts, quotas.

- **Elevation of Privilege:** Tool allow lists, strict argument validation.

Threat	Example in RAG	Primary Controls
Spoofing Tampering	Stolen API keys Poisoned guides, prompt injection	OIDC, short-lived tokens, IP allow lists Signed manifests, allow list, HTML/JS stripping
Repudiation	No trace for risky outputs	Immutable audit log, retrieval IDs, versioned Prompts
Information Disclosure	PII/secret leakage via context	PII redaction, schema filters, access scopes
Denial of Service	Token floods, oversized payloads	Rate limits, circuit breakers, timeouts, quotas
Elevation of Privilege	Unauthorized tool use	Tool allow list, strict argument validation

TABLE 2: STRIDE SUMMARY FOR THE CUSTOMER-SUPPORT RAG SYSTEM.

13. LIMITATIONS AND FUTURE WORK

Failure modes include stale content, ambiguous intents, and out of scope requests. Future work: multimodal inputs (photos/videos), richer local code lookup, and agentic planning for multi step projects. We also plan to integrate confidence scoring and perclaim citations to improve transparency.

14. CONCLUSION

RAG turns fluent conversation into reliable, step-by-step guidance by anchoring responses to enterprise truth.

The result is an approachable "digital apron in the aisle" that improves outcomes for DIYers and pros while reducing operational load.

ACKNOWLEDGMENTS

The author would like to acknowledge that portions of the grammar, phrasing, and linguistic refinement in this manuscript were reviewed and assisted by AI-based language tools to ensure clarity and readability. All technical content, models, algorithms, and research conclusions remain the original work of the author.

APPENDIX: SAFETY POLICY (EXCERPT)

moderation:

pipeline:

- pre_retrieval_filter

- pre_generation_filter

- post_generation_filter

pii:

redact: [email, phone, address, cc_number]

labels:

A1: harassment

A2: hate

A3: self_harm

A4: sexual

A5: unsafe_instruction

A6: privacy_pii

A7: violent_threat

actions:

block: [A2, A3, A5, A7]

redact_and_continue: [A6]

warn_and_continue: [A1, A4]

competitors:

allow_brands: [BrandA, BrandB]

block_brands: [CompX, CompY, CompZ]

domain_allowlist: [example.com, cdn.example.com]

rewrite_on_block: true

security:

auth: oidc

encryption: kms_per_tenant

tool_allowlist: [coverage_calc, inventory_lookup]

REFERENCES

1. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.

2. V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain QA," EMNLP, 2020.

3. S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in IR, 2009.

4. J. Johnson, M. Douze, H. J´egou, "Billion-scale similarity search with GPUs," IEEE TPAMI, 2019.