CAR-RAG: Category-Aware Hybrid Retrieval-Augmented Generation for Hallucination Mitigation

Tatiana Petrova
SnT - SEDAN
University of Luxembourg
Luxembourg, Luxembourg
tatiana.petrova@uni.lu

Dmitrii Koriakov
SnT - SEDAN
University of Luxembourg
Luxembourg, Luxembourg
dmitrii.koriakov@uni.lu

Radu State
SnT - SEDAN
University of Luxembourg
Luxembourg, Luxembourg
radu.state@uni.lu

***Abstract*—We introduce CAR-RAG (Category-Aware hybRid Retrieval-Augmented Generation), an approach to mitigate hallucinations in large language models (LLMs) for real-world deployments. Unlike single-modality pipelines, CAR-RAG conditions retrieval on semantic query categories and adaptively combines vector retrieval with lightweight causal query augmentation. This category-aware routing improves risk profiles by reducing contradicted claims, which are particularly harmful in safety-critical settings.**

We evaluate CAR-RAG on an automotive Q&A dataset comprising over 700 community-provided questions and answers from Stack Exchange's Motor Vehicle Maintenance & Repair forum. The framework achieves approximately 90% factual accuracy and reduces confidently incorrect statements to 6.4%, outperforming dense retrievers and the base LLM in risk-sensitive settings.

These results highlight trade-offs between peak accuracy and robustness and position CAR-RAG as a practical, interpretable, deployment-ready solution for hallucination mitigation. It suits industrial contexts (automotive troubleshooting, service-center support, and technical diagnostics) where high-precision answers are essential. An open-source implementation is available on GitHub,[1].

***Index Terms*—retrieval-augmented generation, hallucination mitigation, hybrid retrieval, factual accuracy, domain-specific question answering**

1. Introduction

Large Language Models (LLMs) achieve strong performance across many NLP tasks but remain prone to hallucinations (unsupported or incorrect statements), which limits their reliability in safety-critical and industrial settings [1], [2]. The issue is salient in both research and deployments, where unreliable outputs can compromise user trust, legal compliance, or operational safety [3]. Recent studies highlight these challenges, including frameworks for detecting knowledge-centric hallucinations [4] and two-tiered encoder-based detection for RAG systems [5]. Mitigation strategies span training-time interventions and inference-time checking, with retrieval-augmented generation (RAG) emerging as a practical way to ground model outputs in external evidence [6].

Recent advances in RAG have focused on optimizing trade-offs between retrieval approaches. [7] contrasted RAG with long-context LLMs and proposed Self-Route for dynamic routing via model self-reflection. [8] investigated RAG workflows, while [9] introduced DSP, a declarative framework compiling LM calls into optimized retrieval pipelines. These developments underscore a trend toward programmatic and adaptive RAG systems deployable at scale.

Modern RAG depends critically on retrieval modality and interaction granularity. Dense retrievers (DPR [10], Contriever [11]) and late-interaction models (ColBERT [12]) advanced open-domain retrieval under BEIR-style benchmarks. HyDE improves zero-shot retrieval via hypothetical document expansion [13]; multimodal RAG has been explored for vision-language tasks [14], [15]. Beyond vector similarity, graph-oriented retrieval has gained attention for relational and causal reasoning, e.g., GraphRAG [16].

No single modality dominates across all question types, motivating adaptive methods aligned with conditional computation and mixture-of-experts routing [17]. Our contribution is CAR-RAG (Category-Aware hybRid Retrieval-Augmented Generation), a framework that conditions retrieval on query categories and fuses vector-based retrieval with lightweight causal query augmentation rather than full graph traversal. CAR-RAG is a practical, interpretable method emphasizing a risk-sensitive objective: minimizing contradicted claims, which are disproportionately harmful in safety-critical applications.

To evaluate factual reliability, we introduce the Factual Accuracy Score (FAS), which penalizes contradicted and unverifiable claims, and the Robustness Across Categories (RAC) criterion, capturing stability across heterogeneous query types. Our experiments compare dense retrieval, graph-oriented retrieval, and CAR-RAG, highlighting trade-offs between peak factual accuracy and risk-sensitive robustness. We apply our approach to the automotive domain and

1. https://github.com/CAR-RAG/CAR-RAG

demonstrate how adaptive retrieval strategies can facilitate reliable LLM applications in industry, with potential for extension to healthcare, legal reasoning, and other high-stakes domains. An open-source implementation is available on GitHub.[2]

2. Problem Statement

We formalize the need for *category-aware* retrieval in RAG: most systems rely on a single modality (typically dense vector search), while graph-based alternatives exist but are less common and more costly [6], [16], [18], [19], [20]. Single-modality pipelines do not explicitly adapt to heterogeneous information needs.

In technical domains such as automotive maintenance and repair, user questions naturally fall into distinct categories:

- **Factual:** requests for specifications or definitions (e.g., *"When do brake rotors really need to be replaced?"*),
- **Causal:** inquiries about mechanisms or reasons (e.g., *"Why are the blades of car radiator fans unevenly spaced?"*),
- **Diagnostic:** troubleshooting scenarios requiring integration of symptoms and causes (e.g., *"How do I test my car battery?"*),
- **Comparative:** evaluations of alternatives across multiple attributes (e.g., *"Aluminium engine vs Cast Iron engine"*).

Retrieval modalities differ across these categories: vector retrieval is effective for factual/comparative queries; causal query augmentation better supports explanatory reasoning; diagnostic queries benefit from a balance of both. Thus a fixed strategy is insufficient for consistently minimizing hallucination. We introduce **CAR-RAG**, a Category-Aware hybRid policy that adaptively combines vector retrieval with causal query augmentation by question category, with an explicit emphasis on mitigating confidently incorrect outputs in risk-sensitive contexts.

Formally, let Q denote a set of user questions, $\mathcal{C} = \{\text{factual}, \text{causal}, \text{diagnostic}, \text{comparative}\}$ the taxonomy of categories, and $\mathcal{M} = \{\text{vector}, \text{causal augmentation}\}$ the retrieval modalities. The task is to define a policy

$$\pi : \mathcal{C} \to \Delta(\mathcal{M}),$$

where $\Delta(\mathcal{M})$ is a probability distribution over modalities, enabling category-specific selection or weighted fusion. For each $q \in Q$, we assign a category label $c(q)$, apply $\pi(c(q))$ to guide retrieval, and generate an answer a_q that minimizes hallucination relative to references.

To evaluate factual reliability, we decompose generated claims into three categories: supported, contradicted, unverifiable. Let $s(a_q)$, $c(a_q)$, and $u(a_q)$ denote their respective counts, with $N(a_q) = s(a_q) + c(a_q) + u(a_q)$. We define a factual accuracy score:

$$\text{FAS}(a_q) = \frac{s(a_q) - \alpha \cdot c(a_q) - \beta \cdot u(a_q)}{N(a_q)},$$

where α and β are penalty weights with $\alpha > \beta \geq 0$. In our evaluation, we set $\alpha = 2.5$ and $\beta = 1$, so that one contradicted claim offsets 2.5 supported claims, while one unverifiable claim offsets 1 supported claim. These values operationalize a safety-oriented perspective: contradicted claims (confidently incorrect statements) represent high-risk errors in industrial deployments, while unverifiable claims primarily reduce utility. We also report contradicted and unverifiable rates separately.

To capture stability across heterogeneous categories, we use the Robustness Across Categories (RAC) metric, defined as

$$\text{RAC} = \gamma \times \text{CS} \; + \; \delta \times \frac{\text{WCP}}{100},$$

where CS is the Consistency Score, WCP the Worst-Case Performance, and γ, δ satisfy $\gamma + \delta = 1$. The Consistency Score is

$$\text{CS} = 1 - \text{CV}, \qquad \text{CV} = \frac{\sigma}{\mu},$$

with μ and σ the mean and standard deviation of category-wise FAS values, and

$$\sigma = \sqrt{\frac{1}{|\mathcal{C}|} \sum_{k \in \mathcal{C}} \left(\text{FAS}^{(k)} - \mu \right)^2}, \qquad \text{WCP} = \min_{k \in \mathcal{C}} \text{FAS}^{(k)}.$$

In our evaluation, we adopt $\gamma = 0.6$ and $\delta = 0.4$, emphasizing uniform reliability (via CS) while maintaining a floor on the weakest category (via WCP).

Finally, a satisfactory solution should provide: (i) **Interpretability**. Transparent, category-dependent retrieval decisions; (ii) **Deployability**. Feasible in API-based environments without full graph construction; (iii) **Domain transferability**. Applicability beyond the automotive domain; (iv) **Risk sensitivity**. Explicit minimization of contradicted errors under practical constraints.

3. Framework Architecture

The Category-Aware hybRid Retrieval-Augmented Generation (CAR-RAG) framework consists of five interconnected stages (Fig. 1): query categorization, category-specific retrieval, adaptive fusion, answer generation, and evaluation.

3.1. Question Categorization

Given a user query q, the system first assigns it to one of the categories $\mathcal{C} = \{\text{factual}, \text{causal}, \text{diagnostic}, \text{comparative}\}$ using a lightweight classifier. In our implementation, GPT-4o-mini was employed for structured claim extraction and verification against gold standard answers from Stack Exchange. Unlike prior work relying on human expert comparison, our evaluation pipeline is fully automated: generated

2. https://github.com/CAR-RAG/CAR-RAG

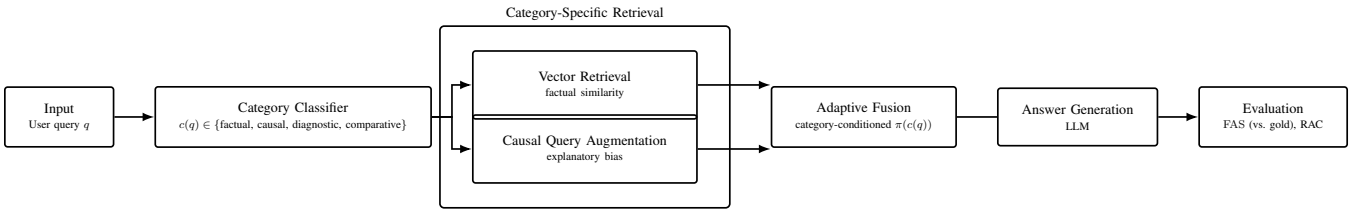


Figure 1. Architecture of the CAR-RAG framework. A user query is categorized into one of four semantic types (factual, causal, diagnostic, comparative). The system routes the query to appropriate retrieval modalities (vector-based factual similarity and causal query augmentation for explanatory emphasis). Retrieved evidence is combined via an adaptive, category-conditioned fusion policy $\pi(c(q))$, guiding the LLM in answer generation. Outputs are evaluated using factual accuracy FAS against gold standard answers and the cross-category robustness metric RAC.

claims are decomposed and checked against authoritative reference answers, enabling factual accuracy assessment without direct human annotation.

Gold Standard Answers. Gold standard answers represent expert-validated, authoritative responses used as the benchmark for evaluation. They were sourced from *Stack Exchange Mechanics* (mechanics.stackexchange.com) under the following selection criteria:

- Marked as "accepted" by the original question askers;
- High community endorsement (median score: 22, range: 2–106);
- Verified contributions from certified automotive mechanics.

These answers are typically long-form (average length $\approx$ 350 words, range 50–1500), technically precise, and practically oriented, often including step-by-step instructions, safety considerations, and causal explanations.

Gold standards provide an objective evaluation baseline: each atomic claim in a generated answer is checked against the reference to determine whether it is supported, contradicted, or unverifiable. This enables systematic classification of hallucination types and reflects professional expertise from experienced mechanics. For example, for the query *"What causes engine knocking?"*, the gold standard answer consists of a detailed expert explanation covering low-octane fuel, carbon buildup, ignition issues, consequences, and diagnostic procedures. Such reference material is critical for accurately distinguishing factual grounding from confidently incorrect (contradicted) or unverifiable claims. It also enables automated factual accuracy assessment.

3.2. Category-Aware Retrieval Policies

Each query category activates specific retrieval modalities:

- **Vector Retrieval:** Dense embedding similarity search using `text-embedding-3-small` accessed via the Tavily API. This modality excels in factual and comparative questions where precise semantic matching is required.
- **Causal Query Augmentation:** A lightweight strategy rather than full graph traversal, which reformulates user queries to emphasize causal terms (e.g., "why", "cause", "reason"). This biases retrieval toward explanatory evidence without requiring explicit graph indices, as further discussed in Section 6.

3.3. Adaptive Fusion Layer

Outputs from the retrieval modalities are combined through category-specific fusion policies $\pi(c(q))$. Fusion weights are heuristically defined based on category-specific rationale: for example, [0.8, 0.2] for factual queries (emphasizing factual grounding), [0.2, 0.8] for causal queries (emphasizing relationships), and [0.5, 0.5] for diagnostic queries (balancing symptoms and causes). These interpretable weights provide a transparent baseline. The framework explicitly supports future extensions with learned weighting via gradient-based optimization or reinforcement learning.

3.4. Answer Generation and Verification

The fused retrieval context is passed to the generation model (GPT-4o-mini), which produces an answer. Generated responses are decomposed into atomic claims and automatically verified against expert-validated ground truth. Evaluation employs two complementary metrics:

- the Factual Accuracy Score (FAS), which accounts for supported, contradicted, and unverifiable claims with penalties $\alpha = 2.5$ and $\beta = 1$, explicitly prioritizing the minimization of contradicted errors;
- the Robustness Across Categories (RAC), which integrates consistency of accuracy across categories and worst-case category performance, reflecting both uniform reliability and resilience to category-specific failures.

3.5. Implementation Considerations

The entire pipeline does not rely on heavy graph indexing or offline preprocessing, which facilitates deployability. At the same time, explicit category conditioning ensures interpretability of retrieval decisions. In practice, this design has been applied to a domain-specific automotive assistant, but the framework is domain-agnostic and supports extension to healthcare, legal, and other high-stakes fields.

From an industrial perspective, the most critical feature of CAR-RAG is its ability to minimize contradicted claims while retaining high factual accuracy, thereby offering a risk-sensitive solution for safety-critical deployments.

4. Results

We evaluate CAR-RAG against three baselines: a Base LLM (no retrieval), a dense retrieval pipeline (Vector RAG), and a graph-oriented retrieval pipeline (Graph RAG). Factuality is reported both as factual accuracy (FAS; share of supported claims) and as the risk-sensitive weighted score FAS with penalties $(\alpha, \beta) = (2.5, 1)$, while stability is summarized by the cross-category robustness metric RAC (§2).

4.1. Dataset Composition

Figure 2 shows the semantic distribution of the evaluation set ($n = 706$) questions with answers. The dataset is imbalanced in a realistic way, with factual and diagnostic queries dominant.

Classifier validation. Category labels were produced by an LLM-based classifier; on a stratified validation sample ($n = 97$) it achieved 90.7% accuracy (95% CI: [84.9%, 96.5%]) and Cohen's $\kappa = 0.852$ (almost perfect agreement), with most errors between diagnostic and factual.

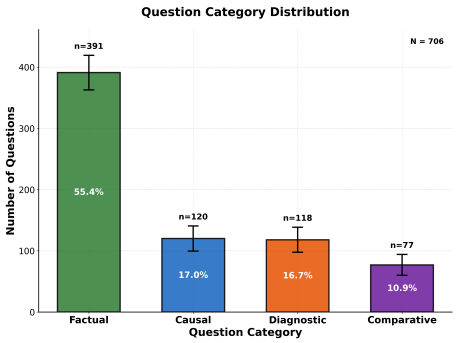


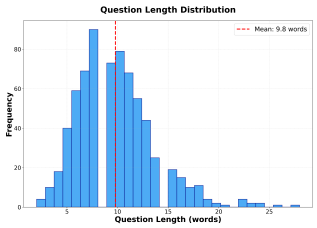
Figure 2. **Question category distribution** ($n = 706$). Factual: 40.5% (286), Diagnostic: 35.6% (251), Causal: 17.0% (120), Comparative: 6.9% (49). Error bars denote 95% confidence intervals (binomial proportion estimates).

Query texts are short while reference (gold) answers are substantially longer (Fig. 3), creating a retrieval challenge for causal and multi-constraint questions.

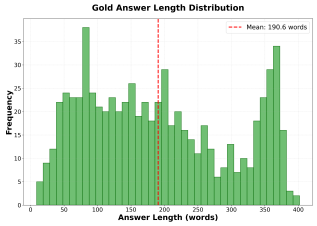
4.2. Overall Performance (Risk-Aware)

Table 1 aggregates the main outcomes: FAS (with 95% CIs) and claim-level error rates. CAR-RAG exhibits the lowest contradicted rate among all methods while maintaining high factuality.

Dense retrieval yields the highest FAS and the fewest unverifiable claims. CAR-RAG achieves the most favorable risk profile by minimizing contradicted claims (confidently incorrect statements), which are disproportionately costly in industrial deployments. In the risk-sensitive metric FAS with $(\alpha, \beta) = (2.5, 1)$, this advantage translates into a competitive overall score and an improved safety margin.



(a) User queries: mean 9.8, median 9.0, range (2, 81).



(b) Gold answers: mean 190.6, median 176.5, range (9, 402).

Figure 3. **Length statistics (questions vs. gold answers;** n=706 **pairs).** Panels report word-level lengths used in automated claim identification (§3).

TABLE 1. **OVERALL OUTCOMES.** FAS (95% CI) AND CLAIM-LEVEL ERROR RATES. FAS IS COMPUTED AT THE ANSWER LEVEL; %CONTRADICTED AND %UNVERIFIABLE ARE COMPUTED AT THE CLAIM LEVEL

Method	FAS (%)	95% CI	Contrad. (%)	Unverif. (%)
Base LLM	73.5	[71.2, 75.8]	15.7	53.6
Vector RAG	**90.1**	[88.4, 91.8]	8.4	**13.6**
Graph RAG	86.3	[84.2, 88.4]	8.6	26.3
CAR-RAG	89.8	[87.9, 91.7]	**6.4**	19.8

4.3. Performance by Category

Figure 4 summarizes per-category factuality (FAS; higher is better). CAR-RAG attains the highest FAS on **Causal** and **Diagnostic** queries, while remaining close to the dense baseline on **Factual** and trailing on **Comparative**. Concretely:

- **Causal:** CAR-RAG **92.0** vs. Vector 91.8 and Graph 86.1.
- **Diagnostic:** CAR-RAG **90.1** vs. Vector 88.2 and Graph 86.1.
- **Factual:** CAR-RAG 89.5 vs. Vector **90.3**.
- **Comparative:** CAR-RAG 87.2 vs. Vector **89.7** and Graph 85.1.

This pattern aligns with the design of CAR-RAG: category-aware fusion particularly benefits causal and diagnostic reasoning, while incurring only a small loss in peak factuality

TABLE 2. **CROSS-CATEGORY ROBUSTNESS (HIGHER IS BETTER).** CAR-RAG: BEST RAC AND BEST WORST-CASE; GRAPH RAG: BEST CONSISTENCY.

Method	Consistency Score (CS)	Worst-Case Performance (WCP, %)	RAC
CAR-RAG	0.983	**88.0**	**0.942**
Vector RAG	0.980	87.7	0.939
Graph RAG	**0.993**	85.2	0.937
Base LLM	0.961	69.3	0.854

on factual/comparative queries. Combined with the lowest contradicted rate (Table 1), this yields a more favorable risk profile for deployment.

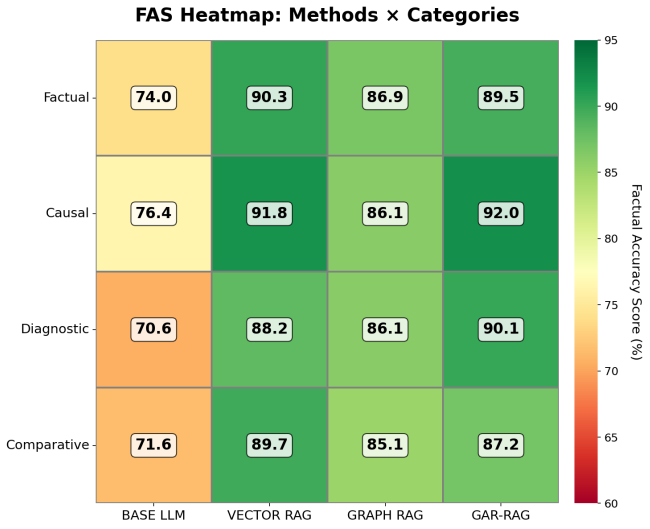


Figure 4. **Per-category factuality (FAS) by method × category.** CAR-RAG is **best on Causal (92.0%)** and **Diagnostic (90.1%)**, numerically close to Vector RAG on Factual (89.5% vs. 90.3%), and below Vector on Comparative (87.2% vs. 89.7%). This reflects the benefit of category-aware fusion on reasoning-heavy queries, while overall maintaining a favorable risk profile (lower contradicted rate; see Table 1).

4.4. Robustness Across Categories (RAC)

We compute RAC as defined in §2 with $(\gamma, \delta) = (0.6, 0.4)$. Using the per-category FAS values in Fig. 4, Table 2 reports Consistency (CS) (higher means less cross-category variance), the Worst-Case category score (higher floor is better), and the composite RAC. CAR-RAG achieves the best overall robustness (RAC=**0.942**) by combining a high Consistency (0.983) with the strongest floor (Worst-Case=**88.0**%). Vector RAG is close (RAC=0.939) with balanced components; Graph RAG yields the highest Consistency (0.993) but a weaker floor (85.2%); the Base LLM trails substantially.

4.5. Error Structure and Hallucination Types

We decompose generated content into Supported, Contradicted, and Unverifiable claims. Figure 5 visualizes the distribution by method. CAR-RAG reduces contradicted claims to the lowest level (6.4%), while Vector RAG attains the fewest unverifiable claims (13.6%).

4.6. Effect of Question Length

Figure 6 plots factuality versus binned question length. Retrieval-based methods show a significant negative correlation between accuracy and length (longer, multi-constraint queries are harder); the Base LLM shows no significant trend. CAR-RAG remains competitive across bins while preserving its favorable risk profile.

Takeaways. (i) Dense retrieval is the strongest single modality by unweighted factuality and exhibits the lowest unverifiable rate; (ii) CAR-RAG achieves a superior risk profile by *minimizing contradicted claims* while remaining competitive in factuality; (iii) Both dense and hybrid pipelines are sensitive to longer, multi-constraint queries, motivating future work on length-aware retrieval and adaptive fusion.

5. Discussion and Theoretical Analysis

Empirical results reveal systematic differences across retrieval modalities under a risk-sensitive objective. CAR-RAG leads on causal and diagnostic queries while maintaining the lowest contradicted rate; Vector RAG is strongest on factual and comparative queries with the fewest unverifiable claims. Cross-category robustness (RAC) is comparable, indicating stable performance with a strong worst-case floor.

5.1. Vector Retrieval and an Information-Theoretic Perspective

Factual and comparative queries often align with compact passages explicitly stating required facts or attributes. From an information-theoretic view, these cases exhibit high pointwise mutual information (PMI) between the query and relevant spans, enabling dense embeddings to retrieve sufficient evidence. However, dense similarity does not penalize contradicted errors; under a cost-sensitive objective such as FAS (with $\alpha > \beta$), even small increases in such claims can dominate the loss.

5.2. Causal Query Augmentation and Relational Structure

Causal and diagnostic questions require reconstructing relationships where structural cues complement lexical alignment. The causal augmentation component reformulates queries toward causal language rather than full graph traversal, biasing retrieval toward explanatory evidence. Alone it does not outperform vector retrieval, but it provides complementary relational coverage valuable in category-wise fusion.

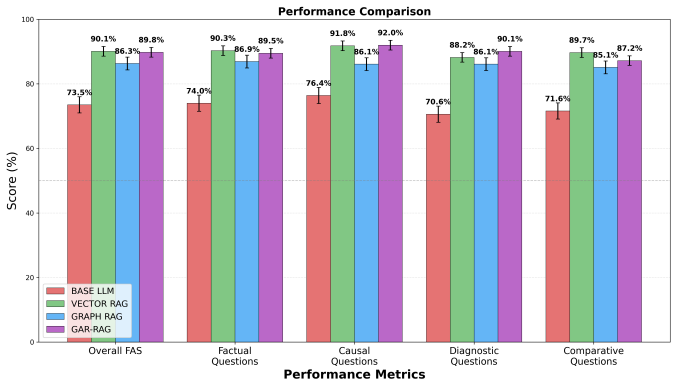


Figure 5. **Claim-type breakdown (supported / contradicted / unverifiable).** Bars show the proportion of *atomic claims* assigned to each label, aggregated over the full test set (claim-level evaluation against gold answers). *Supported* corresponds to unweighted factuality (FAS); in the risk-aware objective FAS with $(\alpha, \beta) = (2.5, 1)$, *contradicted* claims receive the highest penalty. **CAR-RAG** attains the *lowest contradicted rate* (6.4%), yielding the most favorable risk profile; **Vector RAG** achieves the *lowest unverifiable rate* (13.6%) and the highest supported share; **Graph RAG** has contradicted rates comparable to Vector but a higher fraction of unverifiable claims (26.3%); the **Base LLM** exhibits the largest hallucination burden (15.7% contradicted, 53.6% unverifiable). This complements Fig. 4, where CAR-RAG leads on causal/diagnostic queries while trading a small loss in peak factuality for fewer confidently incorrect outputs.

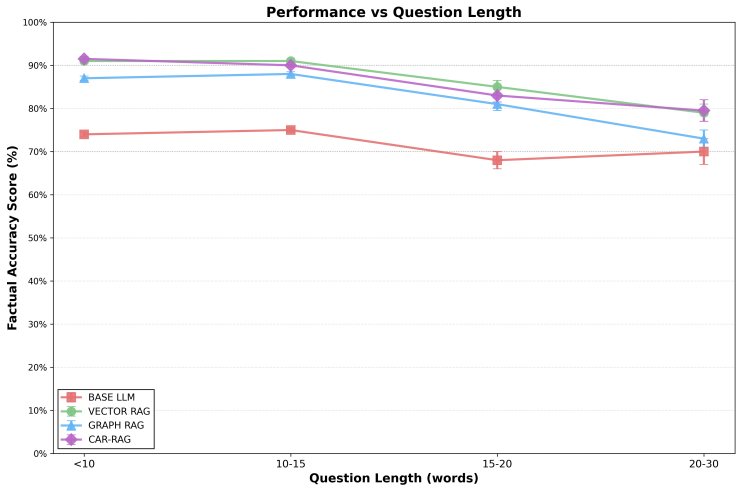


Figure 6. **Factual accuracy vs. question length (mean $\pm$ 95% CI by length bin).** Curves report per-bin means for each method; error bars denote 95% CIs over per-question scores within the bin. Accuracy for retrieval-based methods decreases with length, consistent with increased evidence aggregation needs in longer, multi-constraint queries. Linear associations (Pearson r) between per-question accuracy and length are: *Base LLM* r=0.039 (p=0.28, n.s.), *Vector RAG* r=$-$0.170 (p<0.001), *Graph RAG* r=$-$0.126 (p<0.001), *CAR-RAG* r=$-$0.196 (p<0.001). The decline is most pronounced beyond $\sim$20 words. Despite this trend, **CAR-RAG** preserves competitive factuality (FAS/FAS with $(\alpha, \beta) = (2.5, 1)$) and maintains the lowest contradicted rate across length bins (see Table 1), suggesting a more favorable risk profile for longer and more demanding queries.

5.3. Hybrid Retrieval as a Cost-Sensitive Mixture of Experts

CAR-RAG can be viewed as a mixture-of-experts (MoE) policy where the category label $c(q)$ gates priors over retrieval modalities and induces a cost-sensitive loss via FAS. Two mechanisms drive the gains: (i) **Category gating**—for causal/diagnostic inputs, the policy increases weight on relationally informative evidence, improving coverage and reducing contradicted claims; (ii) **Risk-aware objective**—with $(\alpha, \beta) = (2.5, 1)$, the penalty for contradicted claims dominates, so small reductions yield measurable improvements in FAS. Fixed, interpretable weights ensure strong RAC but may under-adapt within categories; instance-level learned fusion (e.g., differentiable gating or RL with FAS as reward) could reduce this bias and close gaps to the best single-modality scores.

5.4. Length Effects: Coverage vs. Alignment

Accuracy decreases for longer or multi-constraint queries due to a coverage–alignment trade-off: dense similarity favors salient lexical features but may miss low-frequency constraints, whereas relational augmentation improves coverage but can dilute alignment. Long gold answers intensify this, as relevant evidence may be scattered

across passages. Category-aware fusion mitigates contradicted errors by steering retrieval toward causal or diagnostic evidence when needed.

5.5. Cognitive and Practical Considerations

The category-aware design aligns with theories of human information seeking and yields practical benefits: interpretability (explicit category-conditioned routing), deployability (API-only, no heavy graph indices), and a risk-aware error profile (fewer contradicted claims). In safety-critical deployments, this profile is preferable to maximizing unweighted factuality when small gains come at the cost of confidently incorrect statements.

5.6. Summary of Insights

Vector RAG is most effective for factual/comparative queries and yields the lowest unverifiable rate, consistent with an alignment view. Relationally oriented evidence enhances causal/diagnostic reasoning and complements dense similarity without full graph traversal. CAR-RAG combines these strengths, achieving the lowest contradicted rate and near-top robustness; moving from fixed to learned fusion is a promising direction to close remaining gaps across categories.

6. Limitations and Future Work

While CAR-RAG achieves a favorable risk profile (lowest contradicted rate) with competitive factuality and near-top cross-category robustness, several limitations remain and outline concrete directions for future work.

Causal query augmentation as proxy. The current implementation employs a lightweight causal query augmentation strategy rather than full graph traversal, biasing retrieval toward explanatory evidence without explicit graph indices. This proxy cannot capture long-range dependencies or provenance-aware reasoning as in GraphRAG [16]. Future work will explore lightweight graph induction from the corpus (entity/relation extraction with temporal provenance), constrained multi-hop retrieval, graph-aware re-ranking, and staleness detection for time-sensitive facts.

Fixed weighting in CAR-RAG. Modalities are combined with manually defined, category-specific weights (interpretable but static) which may under-adapt within categories and overlook instance-level variation. Promising extensions include learned fusion: (i) differentiable gating (mixture-of-experts) trained to maximize $\mathbb{E}[\mathrm{FAS}_w]$ under $(\alpha, \beta) = (2.5, 1)$; (ii) bandit/RL formulations with risk-sensitive rewards; (iii) Bayesian or meta-learning to personalize weights while retaining interpretability. Joint optimization of FAS and RAC represents another direction.

Category classifier reliability. Category assignment $c(q)$ is produced by an LLM-based classifier that achieved 90.7% accuracy (95% CI: [84.9%, 96.5%]) and Cohen's $\kappa = 0.852$ ("almost perfect") on a validation sample ($n = 97$). Residual misclassification ($\sim$9%), mostly between diagnostic and factual, may propagate to suboptimal fusion. Future work includes expanding the expert-annotated set, adding uncertainty-aware gating or small supervised front-ends, and periodic revalidation across domains.

Evaluation methodology and reference reliability. Automatic claim extraction and verification rely on gold standard answers from Stack Exchange (accepted, expert-validated, high-score responses). Although systematic, this protocol inherits residual noise and may over-penalize paraphrased or alternative-correct answers. Future work will add expert adjudication on a stratified subset, entailment-based (NLI) robustness checks, refined claim segmentation, and partial-credit schemes for equivalent procedures.

Domain scope and transfer. Experiments were limited to automotive maintenance and repair. Generalization to other high-stakes domains (e.g., healthcare, legal, finance) requires validation under domain-specific costs, corpora, and terminology. Future evaluation will target cross-domain and multilingual transfer, and robustness under domain shift and out-of-distribution queries.

Dataset imbalance and fairness across categories. The evaluation set is skewed toward factual and diagnostic queries, which may bias aggregate scores. Future studies will employ balanced or stratified sampling, report per-category intervals alongside RAC, and consider reweighting to reflect stakeholder costs across categories.

Length sensitivity and context constraints. Performance declines for longer, multi-constraint queries due to coverage–alignment trade-offs and context-window limits. Future work includes adaptive top-k retrieval, length-aware chunking and summarization, multi-hop evidence aggregation with redundancy control, and abstention when coverage is insufficient.

Computational and latency considerations. Hybridization introduces retrieval and fusion steps that increase latency and cost. Future research will investigate budgeted retrieval (latency/cost-constrained optimization), passage caching and deduplication, approximate nearest neighbor settings tuned for recall–latency balance, and on-device or edge execution for privacy-sensitive use cases.

Risk sensitivity and metric tuning. The present setup fixes FAS with $(\alpha, \beta) = (2.5, 1)$ and RAC with $(\gamma, \delta) = (0.6, 0.4)$, reflecting industrial preferences (penalize contradicted $>$ unverifiable; favor stability with a floor). Different applications may require alternative trade-offs. Future work will include stakeholder-informed sensitivity analyses and formal selection of $(\alpha, \beta, \gamma, \delta)$ as a multi-criteria decision problem with calibrated confidence and abstention policies.

Evaluation scope and future validation. While the present study focuses on a single-domain dataset (automotive maintenance and repair) and uses fixed fusion weights, the framework and metrics are designed to generalize across domains and evaluation protocols. In future work, we plan to extend CAR-RAG to larger multi-domain corpora and include complementary evaluations with standard retrieval and generation metrics (e.g., EM/F1, Recall@k, calibration error) to benchmark risk-aware measures such as FAS and

RAC against conventional ones. This will help quantify the broader applicability and efficiency–reliability trade-offs for large-scale, data-intensive deployments.

7. Conclusion

We presented CAR-RAG (Category-Aware hybRid Retrieval-Augmented Generation), a framework that combines vector-based retrieval with causal query augmentation (instead of full graph traversal) via interpretable, category-conditioned fusion. The evaluation on 706 automotive maintenance and repair queries employed automated claim extraction and verification against gold standard answers, decomposing outputs into supported, contradicted, and unverifiable claims under a risk-sensitive objective.

Empirically, CAR-RAG delivers a favorable risk profile compared to single-modality baselines. It achieves the lowest contradicted rate (6.4%) while maintaining high factuality (FAS $\approx$ 89.8%) and the best cross-category robustness (RAC $= \mathbf{0.942}$; Vector RAG 0.939, Graph RAG 0.937) with the strongest worst-case floor (WCP $= 88.0\%$, Consistency $= 0.983$). At the category level, CAR-RAG performs best on causal and diagnostic queries, where relational coverage matters most, while dense retrieval remains strongest on factual and comparative queries and yields the lowest unverifiable fraction.

Practically, CAR-RAG satisfies core deployment requirements: (i) *interpretability* through explicit category-conditioned routing; (ii) *deployability* in API-only environments without heavy graph indices; and (iii) *risk-awareness* via metrics that prioritize safety over marginal gains in unweighted accuracy. The key novelty is its explicit focus on minimizing contradicted claims, which impose disproportionate costs in safety- or compliance-critical applications. While CAR-RAG slightly trails dense retrieval on peak factuality for factual and comparative queries, it offers superior robustness and a safer error profile across categories.

Future work will replace fixed fusion with learned, instance-level policies optimized for the risk-sensitive objective, integrate lightweight multi-hop or graph reasoning to improve relational coverage, and validate the framework across additional domains and languages. These steps aim to further reduce contradicted errors while preserving robustness and practicality for real-world deployments.

References

[1] Z. Ji, N. Lee, R. Frieske *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, 2023.

[2] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," in *ACL*, 2022.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2022.

[4] X. Hu, D. Ru, L. Qiu, Q. Guo, T. Zhang, Y. Xu, Y. Luo, P. Liu, Y. Zhang, and Z. Zhang, "Knowledge-centric hallucination detection," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 6953–6975.

[5] I. Zimmerman, J. Tredup, E. Selfridge, and J. Bradley, "Two-tiered encoder-based hallucination detection for retrieval-augmented generation in the wild," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024, pp. 8–22.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Z. Li, C. Li, M. Zhang, Q. Mei, and M. Bendersky, "Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024, pp. 881–893.

[8] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, and X. Huang, "Searching for best practices in retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 716–17 736.

[9] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, and C. Potts, "DSPy: Compiling declarative language model calls into state-of-the-art pipelines," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024, spotlight.

[10] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *EMNLP*, 2020.

[11] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised dense information retrieval with contrastive learning," *Transactions on Machine Learning Research*, 2022.

[12] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over bert," in *SIGIR*, 2020.

[13] L. Gao, Z. Dai, and J. Callan, "Precise zero-shot dense retrieval without relevance labels (hyde)," in *ACL*, 2023.

[14] O. Adjali, O. Ferret, S. Ghannay, and H. Le Borgne, "Multi-level information retrieval augmented generation for knowledge-based visual question answering," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 16 499–16 513.

[15] P. Xia, K. Zhu, H. Li, H. Zhu, Y. Li, G. Li, L. Zhang, and H. Yao, "Rule: Reliable multimodal rag for factuality in medical vision language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 1081–1093.

[16] D. Edge, S. Truitt, J. Larson *et al.*, "A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[17] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, 2022.

[18] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.

[19] G. Izacard and E. Grave, "Unsupervised dense information retrieval with contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[20] K. Santhanam, O. Khattab, N. Rekabsaz, and B. Mitra, "ColBERTv2: Effective and efficient retrieval via lightweight late interaction," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2022, pp. 2241–2251.