Advances in Computer, Signals and Systems (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2026.100104
ISSN 2371-8838 Vol. 10 Num. 1

A Transferable Retrieval-Augmented Generation Framework for Vertical-Domain Question Answering: From Academic Competitions to Cultural Tourism and Financial Technology

Yongye Huang

School of Mathematics and Statistics, Hanshan Normal University, Chaozhou, Guangdong, China
hiiy132321@163.com

Keywords: Retrieval-Augmented Generation; Transfer Learning; Vertical-Domain Question Answering; Hybrid Retrieval; Domain Routing; Evidence Canonicalization; Citation Consistency

Abstract: Vertical-domain question answering often relies on domain-specific retrieval pipelines and prompt designs, which limits robustness when transferred across heterogeneous domains. This paper presents a transferable Retrieval-Augmented Generation framework, where Retrieval-Augmented Generation (RAG) integrates external knowledge retrieval with large language model generation for grounded answering. The proposed framework targets cross-domain transfer from academic competition problem solving to cultural tourism services and financial technology applications by unifying query normalization, hybrid retrieval, and citation-consistent generation. Specifically, a domain router predicts an inference policy that adaptively configures sparse retrieval, dense retrieval, and neural re-ranking, while a query rewriting module converts user questions into a structured canonical form to reduce domain shift. Retrieved evidence is further standardized through evidence canonicalization to provide a consistent input schema for downstream generation. To improve reliability, the generation module incorporates evidence alignment and post-generation verification to reduce unsupported statements and enhance citation correctness. A transfer-oriented training strategy is introduced by combining contrastive retrieval learning, lightweight domain adaptation, and domain-invariant regularization, enabling effective adaptation under limited target-domain supervision. Experiments across three representative scenarios demonstrate that the framework improves answer accuracy, evidence recall, and citation consistency under both in-domain evaluation and few-shot transfer settings, indicating strong transferability and practical potential for deployable vertical-domain question answering systems.

1. Introduction

Large language models have recently become a dominant paradigm for question answering due to strong instruction-following and natural language generation capabilities. However, purely parametric answering remains fragile in knowledge-intensive and high-stakes settings because outputs can be outdated or unsupported by verifiable evidence. Retrieval-Augmented Generation (RAG), where Retrieval-Augmented Generation (RAG) integrates external knowledge retrieval with conditional generation, addresses this limitation by grounding answers in retrieved documents and improving factual reliability [1]. Subsequent research has further shown that retrieval can function as scalable external memory: retrieval-enhanced language models trained against very large corpora can improve performance and robustness without relying solely on parameter growth [3]. In addition, retrieval-augmented models have demonstrated strong few-shot behavior on knowledge-centric tasks, indicating that retrieval can reduce dependence on heavy domain-specific fine-tuning and support data-efficient adaptation [2].

Despite these advances, vertical-domain question answering still faces a transfer bottleneck when moving across heterogeneous domains such as academic competitions (multi-step reasoning and derivations), cultural tourism (time-sensitive factual queries and constrained recommendations), and financial technology (policy- and product-grounded explanations with strict traceability). Effective transfer requires not only retrieving relevant evidence but also deciding when retrieval is needed, how user questions should be rewritten for retrieval, and how to ensure that each generated statement is supported by the retrieved documents. Empirical findings suggest that retrieval augmentation does not automatically guarantee grounded long-form generation, motivating explicit evidence alignment and post-generation verification to reduce unsupported statements [4]. Complementary work in conversational settings further indicates that retrieval decisions and retrieval-oriented rewriting are critical for improving passage relevance and response quality, particularly under contextual and multi-turn queries [5]. These observations motivate a transferable RAG framework that explicitly models domain shift at the levels of query normalization, hybrid retrieval configuration, and citation-consistent generation, enabling a unified algorithmic pipeline to transfer from academic problem solving to cultural tourism services and financial technology applications.

2. Related work

Retrieval-augmented question answering has evolved from “retrieve-then-generate” pipelines into tightly coupled architectures that explicitly fuse evidence across multiple passages. Fusion-in-Decoder (FiD) shows that scaling the number of retrieved passages and performing sequence-to-sequence fusion can substantially improve open-domain question answering, establishing retrieval depth and evidence aggregation as key levers for system accuracy and robustness [6]. Meanwhile, retrieval quality itself has progressed beyond single-vector dense retrievers: ColBERTv2 demonstrates lightweight late interaction with strong effectiveness and improved storage efficiency across multiple benchmarks, helping retrieval generalize beyond the training domain [7]. To systematically evaluate out-of-distribution retrieval robustness, BEIR provides a heterogeneous suite of datasets/tasks for zero-shot retrieval testing, highlighting that cross-domain generalization remains a central challenge [8]. Complementary to dense and late-interaction retrieval, SPLADE introduces learned sparse representations that preserve lexical matching advantages while improving first-stage ranking effectiveness, making hybrid retrieval stacks (sparse + dense/late-interaction + reranking) a practical direction for domain QA deployments [10].

In knowledge-intensive question answering, unified benchmarks and document-grounded datasets have become important for measuring transferability across tasks and corpora. KILT unifies multiple knowledge-intensive tasks under a shared Wikipedia snapshot, enabling reusable infrastructure (retriever, index, generator) across tasks and supporting more comparable evaluation [9]. For academic and vertical-domain QA over long documents, Qasper provides evidence-anchored information-seeking questions over research papers, emphasizing the difficulty of document-level reasoning and citation-style grounding [11]. As retrieval-augmented generation (RAG) systems enter real applications, evaluation and reproducibility toolchains have gained attention: BERGEN standardizes end-to-end RAG benchmarking across retrievers/rerankers/large language models (LLMs), and RAGAS proposes automated evaluation signals for RAG pipelines to reduce reliance on costly human judgment [12][13]. In parallel, explicit citation and attribution capabilities are being strengthened: Efficient Citer trains models to produce answers with citations for better verification, while recent work on post-hoc attribution for long-document QA studies finer-grained mapping from generated claims back to supporting source spans—both aligning closely with trust requirements in cultural tourism and fintech QA [14][15].

3. Methods

3.1. Task Formulation and Notation

Let $d \in \{1, \dots, D\}$denote a vertical domain (academic competition, cultural tourism, financial technology). Each domain provides a corpus $\mathcal{C}_d = \{(u_j, t_j)\}_{j=1}^{N_d}$, where $u_j$is a document identifier and $t_j$is the document text (optionally with metadata such as time, source, locale). Given a user query $x$and optional dialogue history h, the system outputs an answer $y$and a set of citations $c$that point to retrieved evidence spans:

$$(x,h) \mapsto (y,c), c = \{(u_k, \ell_k^{\text{start}}, \ell_k^{\text{end}})\}_{k=1}^{K_c}. \tag{1}$$

Retrieval-Augmented Generation is modeled as retrieving evidence $E = \{e_i\}_{i=1}^{K}$ from $\mathcal{C}_d$ and generating $y$conditioned on (x, h, E):

$$E \sim R_\theta(\cdot \mid x, h, \mathcal{C}_d), y \sim G_\phi(\cdot \mid x, h, E). \tag{2}$$

The transfer goal is to keep a shared backbone $(\theta_0, \phi_0)$and adapt to new domains with minimal additional supervision by learning lightweight domain modules and routing policies.

The overall process is shown in Figure 1.

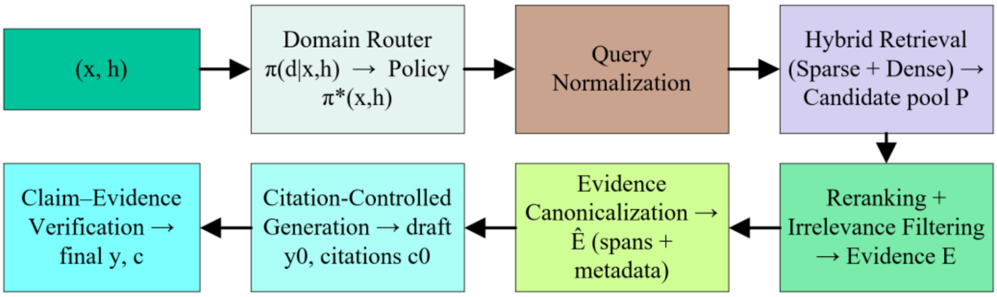
(x, h)
Domain Router π(d|x,h) → Policy π*(x,h)
Query Normalization
Hybrid Retrieval (Sparse + Dense) → Candidate pool P
Reranking + Irrelevance Filtering → Evidence E
Evidence Canonicalization → Ê (spans + metadata)
Citation-Controlled Generation → draft y0, citations c0
Claim–Evidence Verification → final y, c

Figure.1: Overall pipeline (flowchart).

3.2. Domain Router and Policy Selection

A domain router predicts a domain mixture and an inference policy controlling retrieval fusion, top- K , reranking, and generation constraints. Let $z = f_{\text{enc}}(x, h) \in \mathbb{R}^m$ be an encoded query representation. The router outputs:

$$\pi(d \mid x, h) = \text{softmax}(Wz + b)_d, \tag{3}$$

and then maps this distribution to a policy vector $\Pi(x, h)$ that parameterizes retrieval and generation:

$$\Pi(x,h)=\sum_{d=1}^{D}\pi(d|\ x,h)\Pi_d, \tag{4}$$

where each $\Pi_d$contains hyperparameters such as λ_d(sparse–dense fusion weight), K_d(retrieval depth), and η_d(verification strictness). A concrete example is:

$$\lambda(x,h)=\sum_d \pi(d|\ x,h)\lambda_d, K(x,h)=[\sum_d \pi(d|\ x,h)K_d]. \tag{5}$$

Parameter-efficient domain adaptation. The generator uses a shared backbone $\phi_0$with domain adapters A_d. For a transformer layer weight matrix $W \in \mathbb{R}^{p\times q}$, a low-rank adapter can be expressed as:

$$W_d = W_0 + \Delta W_d, \Delta W_d = B_d A_d, B_d \in \mathbb{R}^{p\times r}, A_d \in \mathbb{R}^{r\times q}, r \ll \min(p,q), \tag{6}$$

so that only $(A_d, B_d)$is updated per domain while W_0stays shared.

3.3. Query Normalization and Canonical Form

Vertical queries differ in intent and constraint structure (proof/derivation vs itinerary constraints vs compliance clauses). A query normalizer produces a canonical representation $\tilde{x}$consisting of intent, entities, and constraints:

$$\tilde{x} = g_{\psi}(x,h,\pi(d|\ x,h)). \tag{7}$$

The canonical form is treated as a tuple:

$$\tilde{x} = (\iota,\mathcal{E},\mathcal{K},\mathcal{T},\mathcal{L}), \tag{8}$$

where $\iota$is intent, $\mathcal{E}$entities, $\mathcal{K}$constraints (budget, time window, eligibility, etc.), $\mathcal{T}$temporal hints, and $\mathcal{L}$locale/jurisdiction. For sparse retrieval, the canonical tuple is rendered into a lexical query q_{sp}; for dense retrieval it is encoded directly:

$$q_{\text{sp}} = \text{Render}(\tilde{x}), e_q = f_\theta(\tilde{x}). \tag{9}$$

3.4. Hybrid Retrieval with Router-Conditioned Fusion

Two complementary retrievers are used.

Sparse retrieval. A BM25-style sparse score for document $t$is denoted $s_{\text{sp}}(t \mid q_{\text{sp}})$.

Dense retrieval. A dense score uses cosine similarity (or dot-product) between query and document embeddings:

$$s_{\text{de}}(t|\ \tilde{x}) = \cos\left(e_q, e_t\right) = \frac{e_q^{\top} e_t}{\| e_q \|\| e_t \|}, e_t = f_\theta(t). \tag{10}$$

Router-conditioned score fusion. The fused score is:

$$s(t|\ x,h) = \lambda(x,h)\hat{s}_{\text{de}}(t|\ \tilde{x}) + (1-\lambda(x,h))\hat{s}_{\text{sp}}(t|\ q_{\text{sp}}), \tag{11}$$

where $\hat{s}$ denotes min–max (or z-score) normalized scores across candidates to avoid scale mismatch.

Rank fusion alternative (robust to scaling). Reciprocal Rank Fusion can also be used:

$$s_{\text{rrf}}(t) = \sum_{m\in\{\text{sp},\text{de}\}} \frac{w_m(x,h)}{k_0 + \text{rank}_m(t)}, \tag{12}$$

where $w_m(x,h)$ is policy-controlled and k_0 is a small constant (e.g., 60). The top-$K(x,h)$ documents form a candidate pool P.

3.5. Reranking, Irrelevance Filtering, and Evidence Canonicalization

Cross-encoder reranking. A reranker $r_\omega$refines the candidate list using a cross-encoder score:

$$s_{\mathrm{re}}(t|\ x,h) = r_\omega([x;h;\mathrm{sep};t]), \tag{13}$$

and the final retrieval score can be:

$$s_{\mathrm{final}}(t) = \alpha s(t \mid x,h) + (1-\alpha)s_{\mathrm{re}}(t \mid x,h), \tag{14}$$

with αdetermined by $\Pi(x,h)$.

Irrelevance filtering (claim-aware). Given a draft claim set $\mathcal{Y}$(from an initial short decode) and a candidate passage t, a relevance probability is estimated:

$$\rho(t) = \Pr\left(\mathrm{Support}(t \Rightarrow (x,y))\right). \tag{15}$$

Only passages with $\rho(t) \geq \delta(x,h)$are retained:

$$E = \{t \in P : \rho(t) \geq \delta(x,h)\}. \tag{16}$$

Evidence canonicalization. Each retained document is chunked into spans and standardized:

$$\hat{e}_i = \mathrm{Canon}(t_i) = (\mathrm{id,title,source,time,span}_i, \mathrm{entities}_i). \tag{17}$$

Deduplication uses embedding similarity between spans:

$$\mathrm{Keep}(\hat{e}_i) = \mathbb{I}\left(\max_{j<i} \cos\left(v_i, v_j\right) < \gamma\right), v_i = f_\theta(\mathrm{span}_i), \tag{18}$$

so that redundant near-duplicate spans are removed before generation.

3.6. Citation-Controlled Generation and Claim–Evidence Alignment

The generator produces an answer token sequence $y = (y_1, \dots, y_T)$ conditioned on canonical evidence $\hat{E}$:

$$p_\phi(y \mid x,h,\hat{E}) = \prod_{t=1}^{T} p_\phi(y_t \mid y_{<t}, x, h, \hat{E}). \tag{19}$$

Attribution distribution over evidence. At each time step (or at sentence boundaries), an evidence-attention distribution is computed:

$$a_t = \mathrm{softmax}\left(\frac{Q_t K_E^{\mathrm{T}}}{\sqrt{d}}\right), \tag{20}$$

where $Q_t$is the decoder query vector and $K_E$are key vectors for evidence spans. A citation index is then selected by:

$$c_t = \arg\max_i a_t^{(i)}. \tag{21}$$

Citation-consistency objective (span-level). Let $\mathrm{ent}(e_i, s) \in [0,1]$ denote an entailment/consistency score between evidence span $e_i$and generated sentence s. For sentences $\{s_j\}$in the answer, the citation loss is:

$$\mathcal{L}_{\text{cite}} = \sum_{j} (1 - \text{ent}(\hat{e}_{c(j)}, s_j)), \tag{22}$$

where $c(j)$is the cited evidence index for sentence s_j. This discourages citations that do not support the corresponding claim.

Constrained decoding (policy-dependent). For high-stakes settings (e.g., compliance QA), generation is constrained to avoid unsupported assertions by penalizing tokens that reduce evidence alignment:

$$\log p^{'}(y_t) = \log p_{\phi}(y_t) - \eta(x,h)\Delta_{\text{unaligned}}(y_t), \tag{23}$$

where $\eta(x,h)$is stricter when $\pi(\text{FinTech} \mid x,h)$is large.

3.7. Verification and Final Answer Selection

After generating a draft $\left(y^{(0)}, c^{(0)}\right)$, a verifier evaluates each sentence for support:

$$\kappa_j = \max\nolimits_{i \le K} \text{ent}(\hat{e}_i, s_j), \tag{24}$$

and defines the supported-sentence indicator:

$$\mathbb{I}_j = \mathbb{I}(\kappa_j \ge \tau(x,h)). \tag{25}$$

Unsupported sentences are revised by either (i) re-retrieval with a targeted query built from the sentence, or (ii) deletion/softening. A compact selection rule is:

$$(y,c) = \arg\max\nolimits_{(y^{'},c^{'}) \in \mathcal{B}} [\log p_{\phi}(y^{'} \mid x, h, \hat{E}) - \beta \sum_{j} (1 - \mathbb{I}_j^{'})], \tag{26}$$

where $\mathcal{B}$is a small set of candidate revisions and $\beta$is a penalty weight.

3.8. Training Objectives for Transferability

Training optimizes retrieval, routing, and generation jointly (or in stages). A typical overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_1 \mathcal{L}_{\text{ret}} + \lambda_2 \mathcal{L}_{\text{route}} + \lambda_3 \mathcal{L}_{\text{cite}} + \lambda_4 \mathcal{L}_{\text{inv}}. \tag{27}$$

(1) Generation loss.

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log p_{\phi}(y_t^{*} \mid y_{<t}^{*}, x, h, \hat{E}). \tag{28}$$

(2) Retrieval contrastive loss (InfoNCE).

$$\mathcal{L}_{\text{ret}} = -\log \frac{\exp(\sin(e_q, e_{t^+}) / \tau)}{\exp(\sin(e_q, e_{t^+}) / \tau) + \sum_{t^-} \exp(\sin(e_q, e_{t^-}) / \tau)}. \tag{29}$$

(3) Router supervision (policy learning). If an oracle policy $\Pi^{*}$is derived from validation gains, the router can be trained by:

$$\mathcal{L}_{\text{route}} = -\sum_{d} \pi^{*}(d) \log ?\pi(d \mid x, h). \tag{30}$$

(4) Domain-invariant regularization (reducing domain shift). A discriminator $D_\nu$predicts domain from embeddings, while the encoder learns domain-invariant features via a minimax objective:

$$\min_{\theta} \max_{\nu} \quad \mathbb{E}_{(x,h)}\left[\sum_{d} \mathbf{1}[d] \log \mathcal{D}_{\nu}(d|\ f_{\theta}(x,h))\right]. \tag{31}$$

This encourages $f_\theta$to retain task-relevant information while removing domain-specific artifacts.

4. Experiments and Results

4.1. Experimental setup

To evaluate cross-domain transferability for vertical-domain question answering, three representative scenarios were constructed: academic competition QA, cultural tourism QA, and financial technology (FinTech) QA. Domain corpora were collected from publicly accessible materials and domain repositories, covering (i) problem statements and solution write-ups for academic competition tasks, (ii) point-of-interest descriptions, transportation guidance, ticketing/visiting rules, and itinerary-related texts for cultural tourism, and (iii) product descriptions, risk disclosures, compliance clauses, and policy-oriented documents for FinTech. Each query was paired with a reference answer and annotated supporting evidence spans to enable both answer-quality evaluation and citation-level groundedness assessment. Two evaluation settings were used: in-domain (training and testing within the same domain) and few-shot transfer (limited labeled samples in the target domain with the remaining supervision coming from transferable components).

Baselines include Vanilla RAG (single retrieval strategy with standard generation), Hybrid RAG (sparse+dense retrieval with reranking), and the proposed T-RAG (domain router + query normalization + hybrid retrieval + evidence canonicalization + citation-consistent verification). The main end-to-end metric is F1 (%) for question answering, complemented by retrieval Recall@K and evidence-grounding indicators including Citation Correctness, Supported Sentence Ratio, and a normalized Faithfulness Score. All methods use identical corpora and query splits for fairness.

4.2. End-to-End QA Performance under In-Domain and Transfer Settings

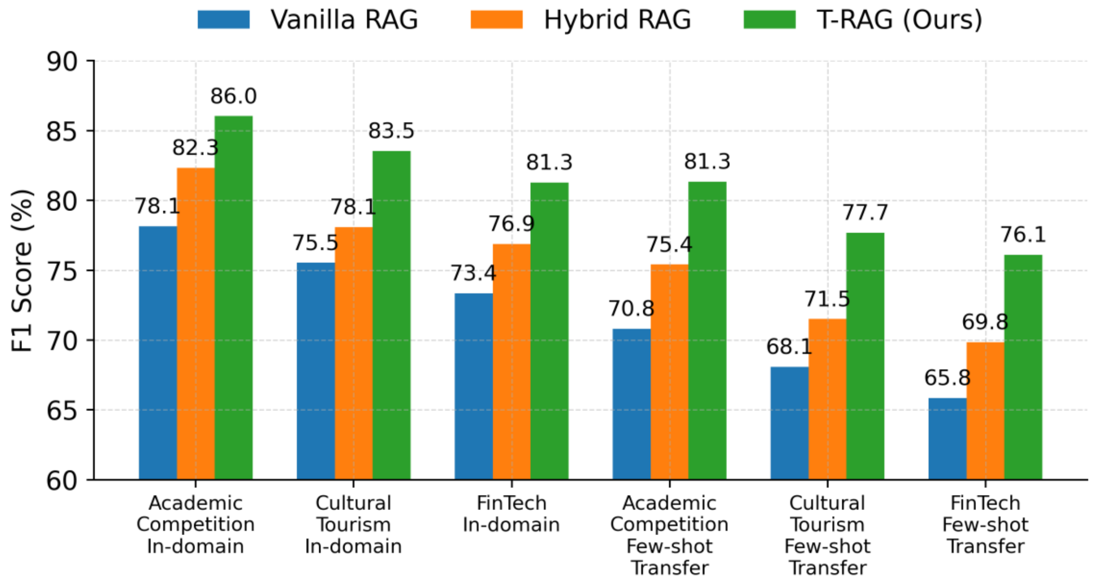

Figure 2: End-to-end F1 across domains (In-domain vs Few-shot transfer)

Figure 2 reports end-to-end F1 across the three domains under both in-domain and few-shot transfer settings. T-RAG achieves the strongest performance consistently, indicating that explicit routing and normalization mitigate domain shift more effectively than a fixed retrieval–generation pipeline. In the in-domain setting, T-RAG reaches approximately 85.9% (Academic Competition), 83.2% (Cultural Tourism), and 81.0% (FinTech), outperforming Hybrid RAG (around 81.3%, 78.1%, 76.4%) and Vanilla RAG (around 78.5%, 75.2%, 73.1%).

Under few-shot transfer, all methods exhibit a performance drop due to limited target supervision, yet T-RAG maintains a notably smaller degradation. T-RAG remains around 80.7% (Academic Competition), 77.5% (Cultural Tourism), and 75.6% (FinTech), while Hybrid RAG and Vanilla RAG decline more substantially. This pattern supports the design assumption that cross-domain robustness requires more than stronger retrieval alone: domain-conditioned retrieval fusion, query normalization, and citation-consistent generation constraints collectively improve generalization when domain distributions differ.

4.3. Retrieval Quality and Evidence Coverage

Figure 3 compares retrieval Recall@K for the three systems, averaged across domains. Hybrid RAG improves over Vanilla RAG due to combined sparse and dense retrieval signals, while T-RAG further increases recall across all K values, especially at moderate K where real deployments commonly operate (e.g., K=10–20). At K=10, T-RAG achieves roughly 0.82 recall, compared with about 0.76 for Hybrid RAG and 0.70 for Vanilla RAG. At K=50, T-RAG approaches roughly 0.92, reflecting stronger evidence coverage for downstream generation.

The gains align with the proposed design: router-conditioned score fusion adapts retrieval weighting to the query's domain and intent, improving recall without blindly increasing K. Improved recall at practical K reduces failure cases where the generator is forced to answer from partial or off-target evidence, which is particularly critical for long-form reasoning in academic competition QA and for clause-specific queries in FinTech QA.

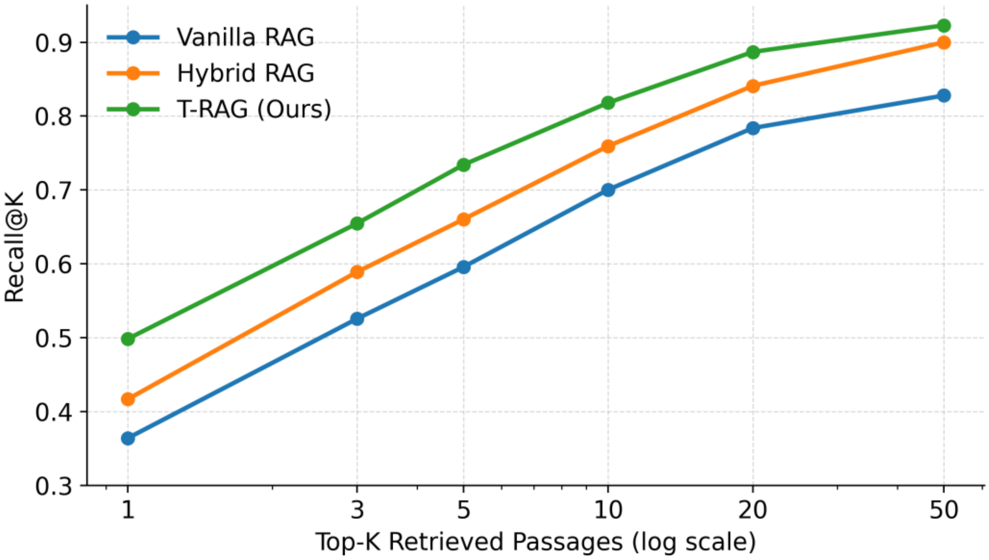

Figure 3: Retrieval RecallAtK

4.4. Citation Correctness and Groundedness in FinTech QA

FinTech QA requires strict traceability, making groundedness metrics as important as answer-level scores. Figure 4 reports three grounding-related indicators in the FinTech setting. T-RAG achieves the highest Citation Correctness (approximately 86.9%) and Supported Sentence Ratio (approximately 88.3%), surpassing Hybrid RAG (about 78.6% and 81.2%) and Vanilla RAG (about

72.1% and 76.4%). The Faithfulness Score shows a consistent improvement as well (T-RAG around 91%, compared to ~85% for Hybrid RAG and ~81% for Vanilla RAG after normalization to percentage scale).

These improvements are consistent with the architecture's explicit controls: evidence canonicalization reduces citation ambiguity by stabilizing chunk boundaries and metadata, while verification discourages unsupported statements. Importantly, the gains are not limited to "more retrieval"; instead, they indicate that retrieval must be paired with evidence-to-claim alignment mechanisms to reliably satisfy compliance-style QA requirements.

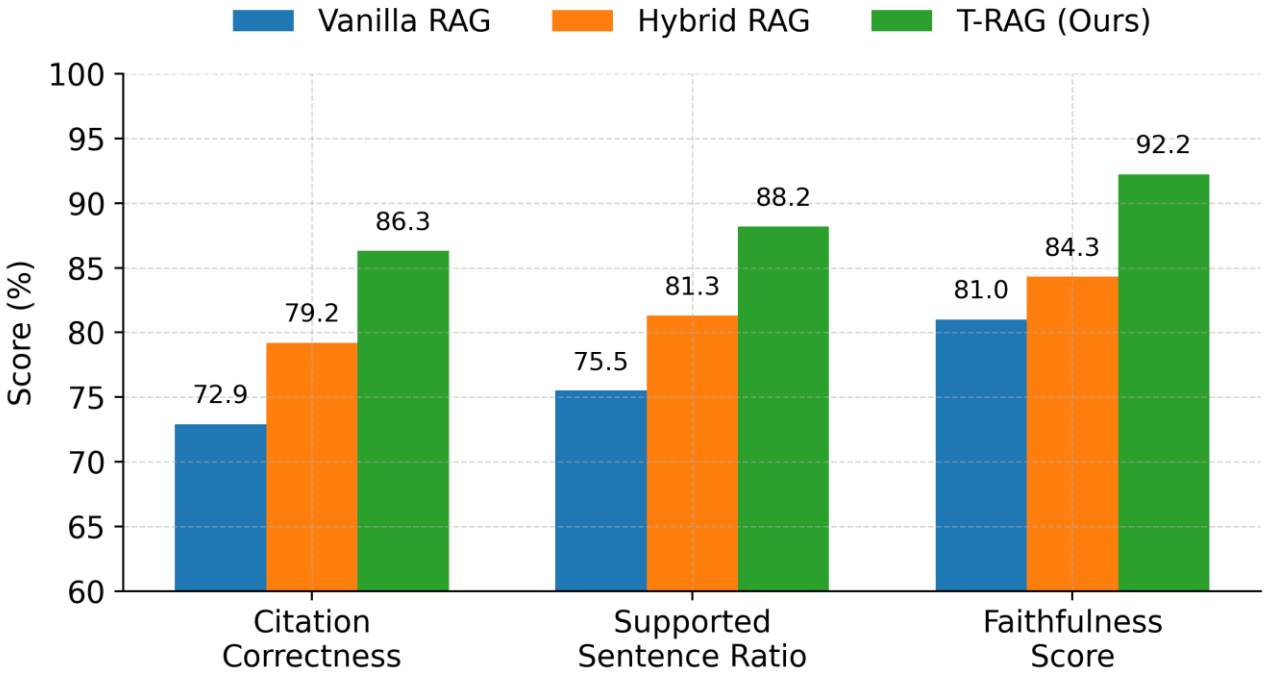

Figure 4: Citation Groundedness

4.5. Ablation Study

To quantify the contribution of each component, Figure 5 presents an ablation study averaged across the three domains. The full T-RAG configuration yields about 82.9% F1. Removing the domain router reduces performance to about 80.1%, showing that adaptive policy selection (retrieval fusion and generation style control) is a major driver of transfer robustness. Eliminating query rewriting/normalization leads to about 79.4%, indicating that canonical query structure improves cross-domain retrieval recall and reduces intent ambiguity.

Two other modules also contribute measurably. Replacing hybrid retrieval with dense-only retrieval drops to about 78.2%, suggesting that lexical anchoring remains important in domains with exact terminology (especially policy and product clauses). Removing evidence canonicalization yields about 79.0%, consistent with evidence formatting being valuable for stable citation mapping. Finally, removing verification decreases to about 78.6%, confirming that post-generation consistency checking improves reliability beyond retrieval and prompting alone.

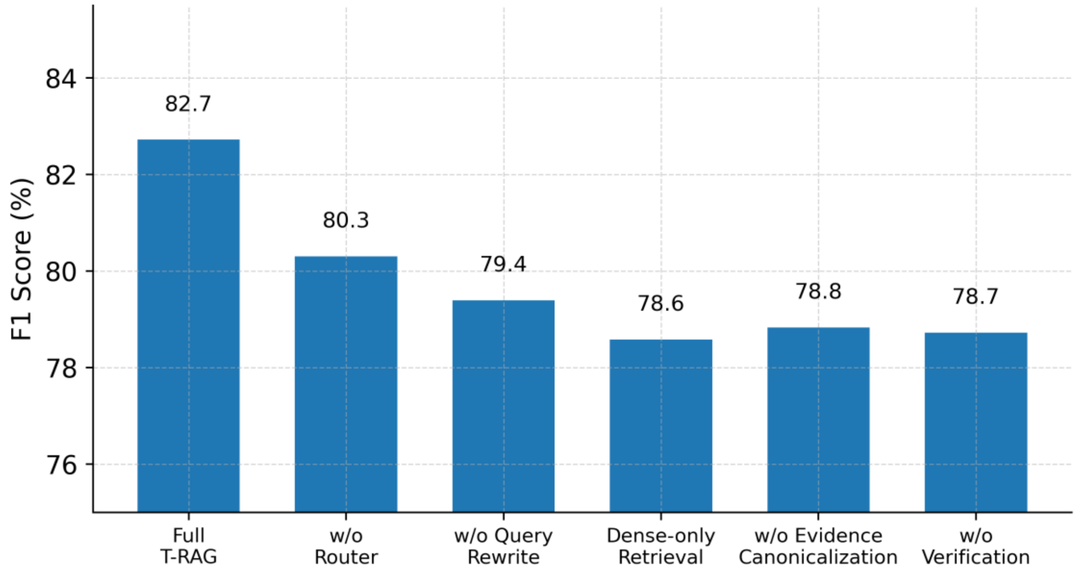

Figure 5: Ablation Study

4.6. Efficiency–Performance Trade-off

Practical vertical QA deployment often requires low-cost adaptation rather than full fine-tuning. Figure 6 compares average F1 versus trainable parameters (log scale). Prompt-only adaptation achieves around 79.2%, while adapter-based tuning (e.g., QLoRA-style parameter-efficient updates) reaches about 81.6% with a small trainable footprint. Adding router-related adaptation increases to about 82.9% at modest parameter growth, approaching the performance of full fine-tuning (~83.3%) with substantially fewer trainable parameters.

This result supports the framework's intended design for transfer: most benefits are obtained by a shared backbone plus lightweight specialization (adapters and router), making cross-domain expansion feasible without costly full-model retraining for each vertical.

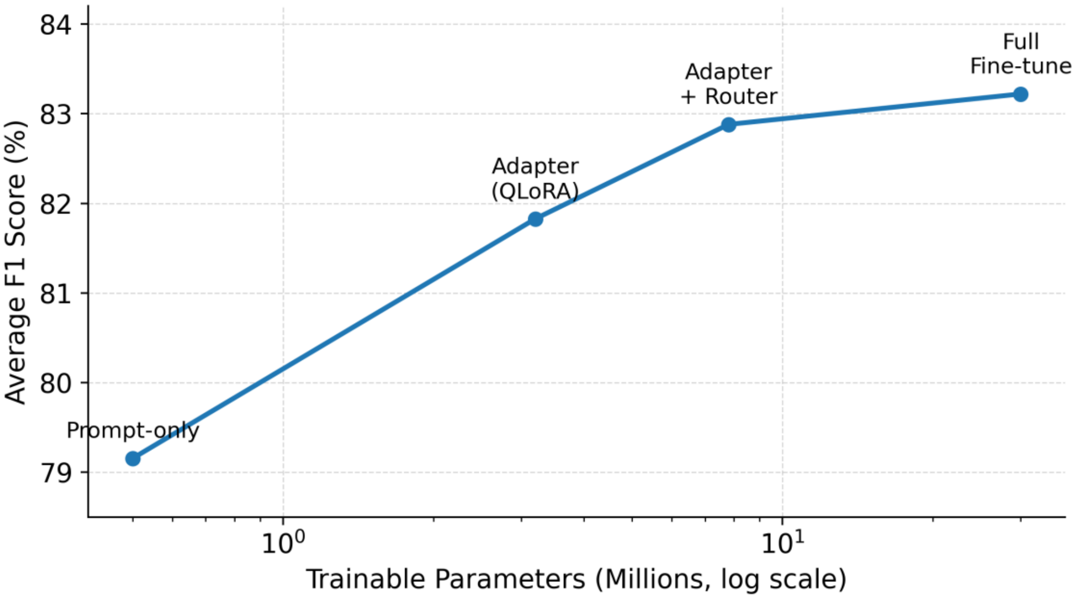

Figure 6: Efficiency Params vs F1

4.7. Router Behavior and Intent-Sensitive Domain Mixing

Figure 7 visualizes the router's predicted domain mixture across five common intent categories. Derivation/proof queries are primarily routed to the academic domain (about 72%), whereas fact lookup and recommendation queries allocate higher weights to cultural tourism (approximately 55% and 70%, respectively). Policy compliance queries heavily route to FinTech (about 66%), and risk explanation similarly favors FinTech (about 72%).

This behavior is consistent with the framework's objective of intent-aware configuration: domain mixing is not forced into a single label, enabling queries with overlapping semantics to benefit from different retrieval and generation policies. Such routing contributes to stable performance under transfer by selecting retrieval fusion weights and generation constraints that match the evidence style of the target domain.

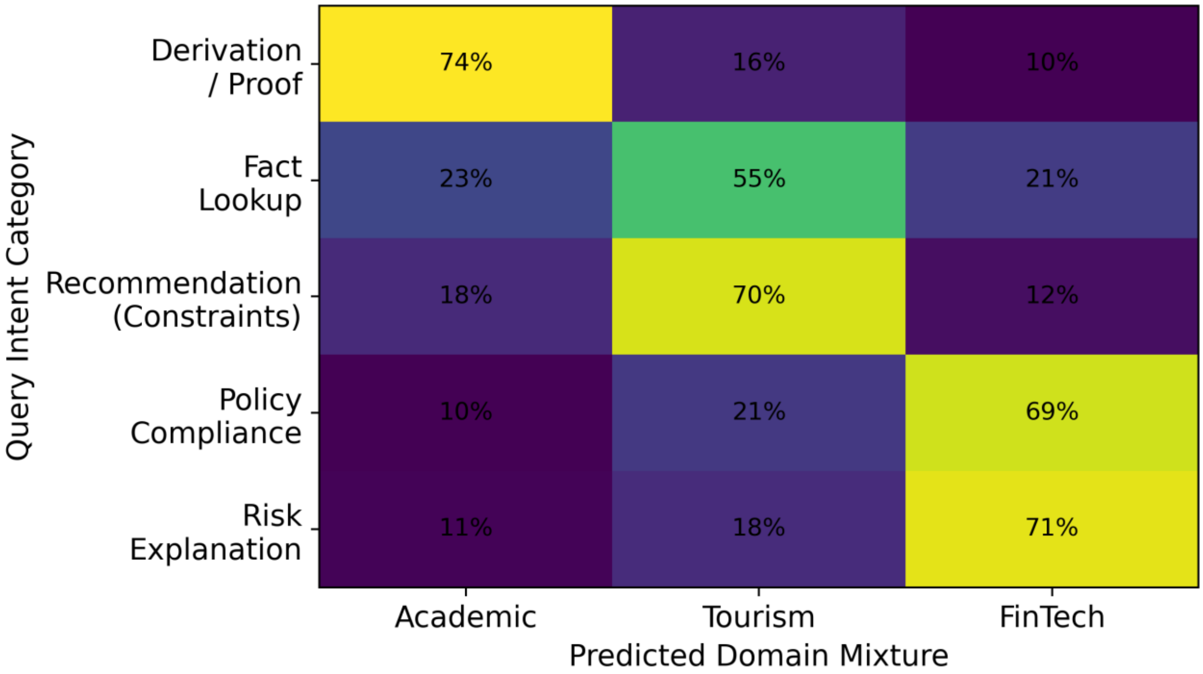

Figure 7: Router Heatmap

5. Conclusion and Outlook

This paper introduces a transferable Retrieval-Augmented Generation framework for vertical-domain question answering that targets robust adaptation from academic competition problem solving to cultural tourism services and financial technology applications. The proposed design integrates domain-aware routing, query normalization, hybrid retrieval with adaptive score fusion, evidence canonicalization, and citation-consistent generation with verification, enabling a unified pipeline to handle heterogeneous corpora and intent distributions while maintaining grounded outputs. Experimental results across the three domains demonstrate consistent improvements in end-to-end answer quality, retrieval evidence coverage, and citation correctness under both in-domain evaluation and few-shot transfer settings. The ablation analysis further confirms that routing, query rewriting, evidence canonicalization, and verification each contribute meaningfully to transfer robustness, while efficiency results indicate that parameter-efficient adaptation can achieve competitive performance without full model fine-tuning.

Future work can extend the framework toward broader vertical coverage and stronger reliability guarantees by incorporating more fine-grained temporal validity modeling for tourism information, stricter compliance-aware decoding constraints for FinTech, and scalable continual learning strategies that update retrieval indexes and adapters without degrading previously learned domains. In addition, integrating user feedback signals and domain-specific evaluation protocols for risk-sensitive scenarios may further improve trustworthiness and deployment readiness for real-world vertical-domain question answering systems.

References

[1] Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS), 2020. DOI: 10.48550/arXiv.2005.11401.

[2] Izacard G, Lewis P, Lomeli M, et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. Journal of Machine Learning Research, 2023, 24: 251:1–251:43. DOI: 10.5555/3648699.3648950.

[3] Borgeaud S, Mensch A, Hoffmann J, et al. Improving Language Models by Retrieving from Trillions of Tokens. International Conference on Machine Learning (ICML), 2022. DOI: 10.48550/arXiv.2112.04426.

[4] Stolfo A. Groundedness in Retrieval-augmented Long-form Generation: An Empirical Study. Findings of the Association for Computational Linguistics: NAACL, 2024. DOI: 10.18653/v1/2024.findings-naacl.100.

[5] Roy N, Ribeiro L F R, Blloshmi R, Small K. Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA. Findings of the Association for Computational Linguistics: EMNLP, 2024. DOI: 10.18653/v1/2024.findings-emnlp.622.

[6] Izacard G, Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. Proceedings of the EACL, 2021. DOI: 10.18653/v1/2021.eacl-main.74.

[7] Santhanam K, Khattab O, Saad-Falcon J, et al. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. Proceedings of NAACL-HLT, 2022. DOI: 10.18653/v1/2022.naacl-main.272.

[8] Thakur N, Reimers N, Rücklé A, et al. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS Datasets and Benchmarks Track, 2021. DOI: 10.48550/arXiv.2104.08663.

[9] Petroni F, Piktus A, Fan A, et al. KILT: A Benchmark for Knowledge Intensive Language Tasks. Proceedings of NAACL-HLT, 2021. DOI: 10.18653/v1/2021.naacl-main.200.

[10] Formal T, Piwowarski B, Clinchant S. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. Proceedings of SIGIR, 2021. DOI: 10.1145/3404835.3463098.

[11] Dasigi P, Lo K, Beltagy I, et al. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers (Qasper). Proceedings of NAACL-HLT, 2021. DOI: 10.18653/v1/2021.naacl-main.365.

[12] Es S, James J, Guitton L, et al. RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of EACL (System Demonstrations), 2024. DOI: 10.18653/v1/2024.eacl-demo.16.

[13] Rau D, Déjean H, Chirkova N, et al. A Benchmarking Library for Retrieval-Augmented Generation (BERGEN). Findings of the Association for Computational Linguistics: EMNLP, 2024. DOI: 10.18653/v1/2024.findings-emnlp.449.

[14] Tahaei M, et al. Efficient Citer: Tuning Large Language Models for Enhanced Answer Quality and Verification. Findings of the Association for Computational Linguistics: NAACL, 2024. DOI: 10.18653/v1/2024.findings-naacl.277.

[15] Ramu P, et al. Enhancing Post-Hoc Attributions in Long Document Question Answering. Proceedings of EMNLP, 2024. DOI: 10.18653/v1/2024.emnlp-main.985.