

A Benchmark Construction and Evaluation Framework for Specialist Domains: Case Study on Defense-Related Documents

Bao Gia Doan[†], Aditya Joshi[†], Pantelis Elinas[†], Aarya Bodhankar[†],
Oscar Leslie[‡], Tom Marchant[‡], Flora D. Salim[†]

[†]UNSW Sydney, Australia

[‡]Cyndr AI, Australia

{bao.doan1, aditya.joshi, p.elinas, a.bodhankar, flora.salim}@unsw.edu.au

{oscar, tom}@cyndr.ai

Abstract

RAG-based question-answering (QA) in specialist domains faces a cold-start problem: no evaluative benchmark and no labeled data for post-training. We present **DoRA (Domain-oriented RAG Assessment)**, a benchmark construction and evaluation framework that needs only a small set of specialist-domain documents, integrating and expert-validating otherwise established components for this cold-start setting. DoRA generates synthetic QA training and evaluation sets with auditable evidence across five domain-specific intents. To mitigate same-pipeline circularity, its training and test splits use *different LLM families* (Claude Sonnet for training; GPT-4o for test) drawn from *disjoint seed-document corpora*. Instantiated on 40 English-language defense documents, DoRA yields ~6.6K curated instances. Against 7 LLM baselines over a 1,259-sample benchmark, a LoRA-adapted Llama-3.1-8B trained on the synthetic split improves all 6 coverage and faithfulness metrics—each gain significant under a paired bootstrap test—and reduces hallucination by more than half under the default GTE retrieval setting, with gains persisting across alternative retrievers, prompting-based baselines, and matched answer-length budgets. Defense experts validate the synthetic QA, calibrate the LLM judge, and supply a fully human-written test set. We position DoRA as a practical framework for specialist-domain RAG under domain shift, with defense as a high-stakes case study.

1 Introduction

Retrieval-augmented generation (RAG) has rapidly become a popular paradigm for knowledge-intensive question-answering (QA) over *domain-specific documents* (Lewis et al., 2020; Gao et al., 2023). Yet RAG evaluation remains dominated by open-domain, Wikipedia-style benchmarks (Kwiatkowski et al., 2019; Yang et al., 2018; Joshi et al., 2017; Petroni et al., 2021). In specialist

domains such as legal, biomedical, defense, and regulated enterprise, three gaps recur: (i) *absence of in-domain evaluation*, since no public benchmark exists for the target domain; (ii) *mismatch of user intents*, since factoid QA under-represents the question types that practitioners ask (Filice et al., 2025; Friel et al., 2024); and (iii) *misaligned metrics*, since precision-oriented overlap scores (e.g., BLEU) reward surface similarity while users need attribution and faithfulness (Zhu et al., 2025). The result is a cold-start problem: performance is hard to measure, supervised data is costly, and retrieval/generation iterations lack sufficient signal to localize the errors. We introduce DoRA (Domain-oriented RAG Assessment), a domain-grounded benchmark construction and evaluation framework that uses a small set of specialist-domain *seed* documents to produce (a) a held-out benchmark split, (b) a larger synthetic split for supervised fine-tuning, and (c) a small expert-curated external check. Our design goals are: (1) **operability on small seed sets** (no reliance on Wikipedia-style coverage), (2) **practitioner-aligned question type coverage** (capture intents beyond factoid QA), and (3) **deployment-relevant, attribution-aware scoring** using modern RAG evaluation protocols (Es et al., 2024; Ming et al., 2024; Chen et al., 2024b; Friel et al., 2024; Zhu et al., 2025). In this paper, we instantiate DoRA on publicly available Defense and National Security material from Australia, using an Alpaca-inspired taxonomy of user intents (Taori et al., 2023): our five styles reduce the instruction intents of open-instruction datasets to the subset answerable from a document corpus (FIND $\approx$ extraction, EXPLAIN $\approx$ open QA, SUMMARIZE $\approx$ summarization, GENERATE $\approx$ open-ended generation, PROVIDE $\approx$ closed attribute extraction), each paired with an evidence budget and grounding checks. We treat Defense as a high-stakes case study that exercises the framework end-to-end, not as a claim that the benchmark is limited to defense

alone. To the best of our knowledge, this is the first synthetic RAG benchmark and evaluation framework specifically instantiated on a defense and national security corpus—prior domain-specific synthetic QA work targets technical manuals (Yuen et al., 2025), enterprise corpora (Filice et al., 2025), or Wikipedia-style multi-hop QA (Chen et al., 2024c; Abaskohi et al., 2025; Wu et al., 2024). The components DoRA builds on—style-conditioned synthetic QA generation, evidence-linked scoring, LLM-as-judge evaluation, and LoRA adaptation—are each established. Our contribution is their *carefully designed integration and validation* for cold-start specialist-domain RAG: a construction design whose circularity controls are built in rather than assumed away, and a validation protocol that lets a domain team certify the result under a small expert-time budget. Concretely:

1. **A domain-grounded benchmark construction and evaluation framework** for RAG-based QA where each QA is paired with auditable evidence bundles, enabling attribution-aware evaluation and reproducible benchmarking in specialized domains where expert time is costly¹.
2. **A lightweight validation protocol for specialized domains without large labeled test sets.** DoRA uses three low-cost checks: a retrieval-conditioned split, an expert-curated evaluation set, and human-LLM-judge calibration on a stratified sample.
3. **A defense-domain case study with benchmark-driven adaptation** on 40 defense documents split into train/test corpora, producing $\sim 6.6\text{K}$ instances. A LoRA-adapted Llama-3.1-8B trained on the cross-generator train split improves RAGEval grounding and reduces hallucination roughly two-fold; the trend persists under alternative retrieval and expert-curated evaluation.
4. **Manual evaluation by defense-domain experts** to evaluate dataset quality, LLM-as-judge reliability and generalization to a completely human-written dataset.

2 Related Work

QA benchmarks (Petroni et al., 2021; Kwiatkowski et al., 2019; Joshi et al., 2017; Rajpurkar et al.,

¹Code, prompts, seed-document sources, derived benchmark instances with provenance fields, and the LoRA adapter are released at <https://github.com/unsw-nlporg/DoRA>; per-component licenses are itemized in Appendix N.

Work	S	E	MM	Seed	Dom
Few-shot Multi-hop QA (Chen et al., 2024c)	✓	✗	✗	✓	Open
FM2DS (Abaskohi et al., 2025)	✓	✗	✓	✓	Open
SMMQG (Wu et al., 2024)	✓	✗	✓	✓	Open
Prompting-based SDG (Schmidt et al., 2024)	✓	✗	✗	✗	Open
RAGAS (Es et al., 2024)	✗	✓	✗	N/A	Open
ARES (Saad-Falcon et al., 2024)	✗	✓	✗	N/A	Open
RAGChecker (Ru et al., 2024)	✗	✓	✗	N/A	Open
RAGEval (Zhu et al., 2025)	✓	✓	✗	✗	Scenario
KIQA Gen. (Yuen et al., 2025)	✓	✗	✗	✓	Domain
DataMorgana (Filice et al., 2025)	✓	✗	✗	✓	Domain
DoRA (Ours)	✓	✓	✗	✓	Domain

Table 1: Comparison of synthetic QA generation and RAG evaluation. S=Synthetic, E=RAG Evaluation, MM=Multimodal, Seed=Seed-doc grounded, Dom=Context scope (Open / Scenario / Domain).

2016) are largely open-domain and Wikipedia-centric, while retrieval benchmarks such as BEIR (Thakur et al., 2021) and multi-hop QA testbeds (Tang and Yang, 2024) target retriever generalization or compositional reasoning rather than deployment over a new private corpus. For RAG evaluation, recent frameworks including RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024), FaithEval (Ming et al., 2024), LLM-in-RAG benchmarking (Chen et al., 2024b), RAGBench (Friel et al., 2024), and RAGChecker (Ru et al., 2024) provide useful diagnostics for faithfulness, attribution, and hallucination. Their common assumption, however, is that a suitable evaluation set already exists. Closer to DoRA are approaches that synthesize or structure evaluation for specific domains. RAGEval (Zhu et al., 2025) uses scenario templates and synthetic articles for scenario-controlled testing; KIQA (Yuen et al., 2025) generates questions from technical manuals; and DataMorgana (Filice et al., 2025) synthesizes Q&A over private enterprise corpora with user-defined question categories. DoRA’s advance over this line of work is not a new component but the way three existing ones are integrated and jointly validated in a single framework: synthetic benchmark construction from authentic seed documents, evidence-linked scoring for grounding and attribution, and a lightweight validation protocol with an expert-answered external check. Each element has precedent in isolation; what prior work lacks is an assembly in which the construction design itself controls circularity and the resulting benchmark is certified by domain experts under a realistic expert-time budget.

Synthetic pipelines such as DoRA reduce anno-

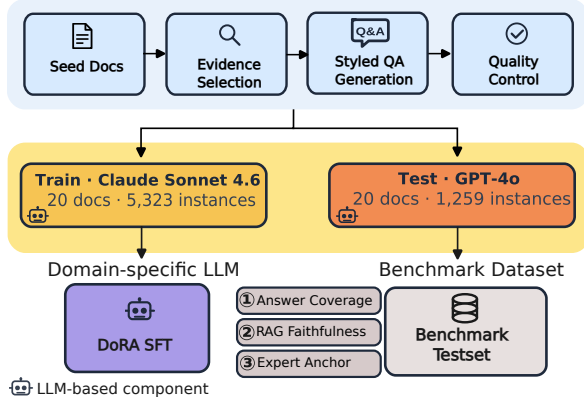


Figure 1: DoRA pipeline, from seed documents to grounded, style-conditioned QA generation, and downstream domain evaluation and adaptation.

tation costs and enable controlled stress tests; related approaches include zero/few-shot prompting for QA generation (Ye et al., 2022; Schmidt et al., 2024; Puri et al., 2020), few-shot multi-hop synthesis over Wikipedia pairs (Chen et al., 2024c), multi-modal multi-hop synthesis (Abaskohi et al., 2025), and modality/style-conditioned question generation (Wu et al., 2024). Against open-domain QA and retrieval benchmarks (Kwiatkowski et al., 2019; Yang et al., 2018; Joshi et al., 2017; Petroni et al., 2021; Thakur et al., 2021) and existing RAG evaluation frameworks (Es et al., 2024; Saad-Falcon et al., 2024; Ming et al., 2024; Chen et al., 2024b; Friel et al., 2024; Ru et al., 2024), DoRA adds three things that matter in specialized domains: grounding in authentic seed documents, coverage of five practitioner-aligned intent styles, and attribution-aware scoring paired with a small external validation protocol. A comparison with existing work is summarized in Tables 1 and 15.

3 Methodology

Let $\mathcal{C}$ denote a domain corpus segmented into passages with IDs, and let $\mathcal{S}$ be the five intent styles (FIND, EXPLAIN, SUMMARIZE, GENERATE, PROVIDE). DoRA produces $\mathcal{D} = \{(q_i, a_i, s_i, E_i)\}_{i=1}^N$, where each instance is a quadruple (q, a, s, E) comprising a question q , reference answer a , style $s \in \mathcal{S}$, and supporting passages $E \subset \mathcal{C}$. An LLM-based agent produces a prediction $\hat{a}$ for each question q . In RAG mode, we model this as: $\hat{a} = G(q, R_k(q))$ where R_k is a retriever that returns the top- k passages from $\mathcal{C}$ and G is a generator that conditions on both the question and retrieved context (with the LLM-only setting as the

special case $R_k(q) = \emptyset$). The pipeline of DoRA is in Figure 1.

3.1 Data Preparation

The starting point is a *minimal* input set: (i) a set of seed documents from the target domain, (ii) a short style specification (we use five intent styles prescribed by defense experts in this context: FIND, EXPLAIN, SUMMARIZE, GENERATE, PROVIDE), and (iii) a bounded amount of expert time, reserved mainly for external validation rather than bulk labeling. In our case study, the seed set is 40 defense documents and the expert-time budget is small (Appendix B). This matters because the framework is intended for settings where a large labeled corpus is unavailable and expert annotation time must be utilized strategically.

We construct DoRA from 40 *seed documents* spanning defense strategy, capability, procurement, and industry-planning reports. Each document is chunked into passages, forming the corpus $\mathcal{C}$ used for evidence selection and retrieval.

3.2 Synthetic Data Generation

The dataset is created using three steps: first select a candidate evidence E per seed, generate QA pairs conditioned on style and finally, perform quality control. For each style s , we select a small evidence set E from the current seed and prompt a style-specific LLM template (see Appendix O) to generate (q, a) conditioned on (s, E) ; we store E for auditability. Each style has an evidence budget k_s , which shapes evidence selection and downstream quality checks. We formulate the synthesis step as sampling from a *style- and evidence-conditioned* generator: $(q, a) \sim p_\phi(\cdot | s, E)$, where ϕ denotes the parameters of the chosen LLM and prompt configuration (chosen parameters in Appendix H).

(Task-conditioned) Evidence selection: Evidence selection chooses a small evidence bundle (reference contexts) E conditioned on the target style s , which determines both the desired cardinality $k_s \triangleq |E|$ and the selection behavior (e.g., extractive single-chunk evidence for FIND versus multi-chunk synthesis for long-form styles). During dataset construction, the candidate pool is typically restricted to passages from the current seed document, $\mathcal{C}_{\text{seed}} \subset \mathcal{C}$. We first apply a Natural Language Inference (NLI)-based prefilter to remove clearly irrelevant candidates and keep a reduced candidate pool $\tilde{\mathcal{C}}_{\text{seed}}$ (top- K_s by NLI relevance). We then use a local LLM scorer to rank candi-

Style	#	Avg $ q $	Avg $ a $	Refs	Qual.
EXPLAIN	165	13.67	33.59	2.99	0.78
FIND	421	12.65	20.83	1.00	0.88
GENERATE	309	13.76	31.42	3.34	0.78
PROVIDE	155	14.00	30.45	1.91	0.76
SUMMARIZE	209	12.85	29.35	2.65	0.82
Total/Avg	1259	13.26	27.70	2.22	0.82

Table 2: DoRA benchmark statistics by style: sample counts, average question/answer lengths (in tokens), average references per item, and average construction-time quality score. **Qual.** is the mean composite score $Q_s \in [0, 1]$ of *accepted* items—a construction-time filter statistic, not a benchmark-time metric.

date bundles drawn from $\tilde{C}_{\text{seed}}$. This scorer explicitly rates *completeness*, *complementarity* (non-redundant coverage), *coherence*, and *task fitness*; the highest-scoring bundle is returned. Formally, we select:

$$E^* = \arg \max_{\substack{E \subseteq \tilde{C}_{\text{seed}} \\ |E|=k_s}} g_s(E), \quad (1)$$

where $g_s(E)$ is a style-conditioned *bundle* scorer computed from the rubric dimensions above (with configurable weights and bounded search over candidate bundles). Details are in Appendix H.

Grounding and Quality Control: Each instance stores an auditable evidence bundle E (passage references), and we accept a generated item only if it satisfies *style-specific* deployability constraints and meets an *evidence-grounded* quality threshold:

$$\begin{aligned} \text{accept}(q, a, s, E) &\iff h_s(q, a, E) = 1 \\ &\quad \wedge Q_s(q, a, E) \geq \tau_s, \end{aligned}$$

where h_s encodes style-specific hard checks (e.g., non-empty fields; PROVIDE must contain numeric content) and Q_s is a composite, evidence-grounded score with threshold τ_s .

We evaluate grounding with respect to E using evidence-based metrics, including sentence-level natural language inference (NLI) with penalties for contradictions, extractive span matching (SQuAD-style span F1) (Rajpurkar et al., 2016), and checks for numerical and date consistency when applicable. For longer-form intents (namely, EXPLAIN, GENERATE and SUMMARIZE), we additionally compute attribution-focused coverage (context recall and precision relative to E) and question–answer relevance (semantic similarity between q and a), in line with standard RAG evaluation protocols (e.g., Es et al., 2024). Specifically, Q_s com-

bines a style-dependent set of metrics $\mathcal{M}_s$ using configurable weights $w_{s,m}$:

$$Q_s(q, a, E) = \sum_{m \in \mathcal{M}_s} w_{s,m} m(q, a, E), \quad (2)$$

after which we apply deduplication and diversity filtering to eliminate repeated patterns and ensure balanced representation across the five styles. The weights $w_{s,m}$ are style-dependent and encode what each intent must get right: the extractive styles (FIND, PROVIDE) concentrate mass on entailment and span alignment, whereas the synthesis styles concentrate it on keypoint recall and question–answer relevance. Concrete weights, thresholds, and the over-generation yield are in Appendix I. This procedure yields our DoRA dataset, composed of the quadruples described above.

3.3 Construction-time vs. benchmark-time evaluation

A run of DoRA on two *disjoint* seed document sets yields two primary artifacts: (i) a held-out synthetic evaluation split for benchmarking and (ii) a cross-generator, cross-corpus training split for supervised fine-tuning; the evaluation corpus can additionally support an expert-answered set at modest expert cost. The quality controls above are used only to *construct* the benchmark. Once the dataset is fixed, benchmark-time evaluation runs independently, with *held-out* questions, fixed retrievers or oracle evidence, and a separate three-layer metric stack (Section 4), so DoRA remains a stable harness as corpora, retrievers, prompts, and generators change.

3.4 Note on cross-generator, cross-corpus construction

A standard pitfall in synthetic benchmarks is that the same LLM and the same source documents produce both the training supervision and the evaluation references, creating a train/test correlation that confounds genuine domain adaptation with pipeline-specific style mimicry. DoRA addresses this on two independent axes at once: the training split is generated by **Anthropic Claude Sonnet 4.6** and the test split by **OpenAI GPT-4o**, both under the same style templates and quality controls (*cross-generator separation*); and the splits are drawn from *disjoint* seed documents with no document-level overlap, verified by passage-level n-gram and embedding-similarity checks (*cross-corpus separation*). Construction-time judges come from a

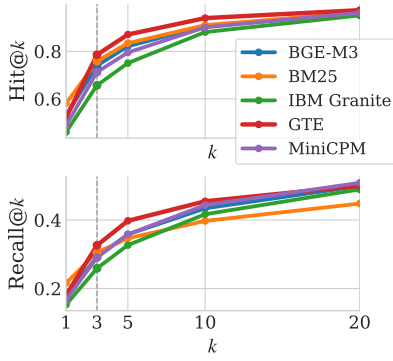


Figure 2: Retriever performance across top- k on DoRA test set.

third family (**Google Gemma**) to limit family-level coupling. A model fine-tuned on the training split therefore cannot win on the test split by imitating GPT-4o’s style or by memorizing test-corpus content. One direct consequence is that train and test reference answers have systematically different length distributions (Claude answers are longer), which shapes how token-overlap metrics should be read; we return to this in Section 4.

4 Experiment Setup

Data and splits. We partition the 40 publicly available defense documents (authored in English, for readers whose primary language is the English of the target country) into two disjoint 20-document corpora: one used *only* for training-split generation, the other *only* for test-split generation. The held-out test split (1,259 instances, GPT-4o-generated; Table 2) and the training split (5,323 instances, Claude Sonnet 4.6-generated, 90/10 train/eval; Appendix F) are disjoint at both document and instance levels. We evaluate retriever performance (Hit@ k , Recall@ k ; see Appendix D) and end-to-end QA under two retriever conditions: a fixed dense retriever (default GTE-multilingual (Zhang et al., 2024); IBM Granite (Awasthy et al., 2025) as a robustness check) and an *oracle* retriever that returns the stored gold evidence bundle E to isolate generation from retrieval error.

Why standard QA metrics are insufficient. The cross-generator construction (Section 3.4) induces systematic length asymmetry between Claude-generated training references, GPT-4o-generated test answer references, and SFT-model outputs. In this regime, precision-oriented overlap scores (Token F1, BLEU, ROUGE-L F1) penalize verbose-but-correct answers, learned sentence-

pair scores absorb length and reference-style cues, and *none* of these reference-comparison metrics measure whether the answer is grounded in the retrieved evidence—the property a deployed RAG system is judged on.

Three-layer metric stack. We therefore adopt a three-layer evaluation in which each layer compensates for a different weakness of the others. (i) *Answer coverage*: recall-oriented variants of three established metrics spanning lexical to semantic granularity—Token Recall, ROUGE-L Recall (Lin, 2004), and BERTScore Recall (Zhang et al., 2019)—plus Length Ratio (generated/gold answer word-count) as a transparency control. All three are length-fair to verbose grounded answers: extra generated tokens cannot lower the score so long as the gold content is covered. (ii) *Grounding at scale*: the RAGEval rubric judge (Zhu et al., 2025), which extracts gold-answer keypoints and labels each as covered correctly (*completeness*), addressed incorrectly (*hallucination*), or unaddressed (*irrelevance*)—the only practical keypoint-level attribution signal at 1,259-item scale (Appendix E). (iii) *Expert ground-truth anchor*: a **50-item stratified human evaluation** (Section 5.4) and an **independently authored 86-item expert-curated set** (Appendix B), used to calibrate the judge rather than scaled to the benchmark. Model identifiers, decoding settings, and per-style evidence budgets are in Tables 19 and 20.

5 Results

5.1 Retriever Evaluation

Following RAGEval (Zhu et al., 2025), we evaluate four closely related retrievers, along with IBM Granite retriever (Awasthy et al., 2025), across various top- k values on the DoRA benchmark test set. Figure 2 shows that, on defense documents, BM25 provides strong early hits but lags dense retrievers in broader coverage at larger k , motivating dense/hybrid retrieval as the natural default for deployment-facing RAG. Overall, GTE achieves the strongest balance across Hit/Recall@ k on the DoRA benchmark set, so we adopt it as the default retriever for end-to-end QA below.

5.2 End-to-End QA and DoRA SFT

We evaluate end-to-end QA with a fixed retriever (GTE-multilingual-Base) at $k=3$, a modest-context deployment setting that bounds the generator input budget while still surfacing multi-passages evidence.

Model	Len. Ratio	Answer Coverage (Recall)			RAG Faithfulness (LLM judge)		
		ROUGE-L-R↑	Tok-R↑	BERTScore-R↑	Comp.↑	Hall.↓	Irrel.↓
GPT-4o-mini	3.44±0.06	67.08±0.11	67.41±0.09	89.00±0.04	68.72±0.65	4.71±0.15	26.57±0.51
GPT-3.5-Turbo	2.05±0.01	64.00±0.14	64.88±0.06	89.70±0.02	65.70±0.12	3.64±0.30	30.66±0.33
Llama-3.1-8B-Instruct (Base)	2.97±0.04	63.30±0.80	62.76±0.74	88.01±0.20	66.72±0.51	4.75±0.23	28.53±0.33
Llama-3-8B-Instruct	3.40±0.18	65.98±1.42	65.82±1.65	88.56±0.39	66.16±0.54	4.53±0.54	29.31±0.09
GPT-4o	2.45±0.05	65.89±0.05	66.09±0.15	89.14±0.05	69.40±0.16	4.93±0.47	25.67±0.32
Minstral-8B-Instruct	5.00±0.01	65.92±0.02	58.32±0.17	86.79±0.02	68.89±0.66	4.49±0.17	26.62±0.55
Qwen2.5-7B-Instruct	3.14±0.05	66.26±0.19	66.60±0.21	89.26±0.06	66.87±0.14	4.53±0.18	28.60±0.18
DoRA SFT	5.09±0.08	69.12±0.19	71.13±0.16	89.11±0.07	73.37±0.40	2.17±0.35	24.47±0.52
<i>Gain vs Base</i>		+5.82	+8.37	+1.10	+6.65	+2.58	+4.06

Table 3: End-to-end QA with GTE retrieval (1,259 test samples, top- $k=3$). **Answer Coverage** reports recall-oriented variants spanning lexical to semantic granularity (Token Recall, ROUGE-L Recall, BERTScore Recall). **RAG Faithfulness** reports RAGEval keypoint-level completeness/hallucination/irrelevance under an LLM rubric judge. Len. Ratio is generated/gold answer word-count ratio (transparency control). All metrics are % except Len. Ratio. Every entry is mean±std over three runs: three fine-tuning seeds for DoRA SFT and three inference runs for each baseline. Gain vs Base is the difference of means against Llama-3.1-8B-Instruct (base); all six gains are statistically significant (Table 4). **Bold/italic** mark best/second-best. ↑ higher is better; ↓ lower is better.

Following Section 4 we report the recall-oriented coverage metrics alongside the RAGEval grounding diagnostics; the same metric set under oracle and IBM Granite retrieval is in Appendices A.1 and A.2. Table 3 shows that, under the same retrieved context, larger off-the-shelf LLMs remain competitive on individual semantic or irrelevance metrics, but no baseline dominates across the coverage and grounding diagnostics—leaving room for domain adaptation.

DoRA SFT. We perform supervised fine-tuning (SFT) on an open model (Llama-3.1-8B-Instruct)² with LoRA (Hu et al., 2021) on the cross-generator DoRA training split (5,323 Claude Sonnet 4.6 instances from the 20-document train-only corpus, 90/10 train/eval). Each example concatenates the question with evidence context $\text{ctx}(E)$ and predicts the reference answer, mirroring end-to-end RAG. We optimize:

$$\min_{\theta} \sum_{(q,a,E) \in \mathcal{D}_{\text{train}}} -\log p_{\theta}(a \mid q, \text{ctx}(E)), \quad (3)$$

where θ denotes the (adapter) parameters. Table 3 shows that the cross-generator-trained DoRA SFT is stable across three fine-tuning seeds and improves over the base Llama-3.1-8B-Instruct most clearly on grounding: mean hallucination falls from 4.75% to 2.17%, with parallel gains in RAGEval Completeness and Irrelevance. Although DoRA SFT does not top every metric—GPT-3.5-Turbo attains the highest BERTScore Recall by answering

²We choose Llama 3.1-8B for its deployment-friendly footprint and prior adoption in defense and national security AI contexts (Meta AI, 2024; Thomas, 2024).

tersely, which suits a metric computed against short GPT-4o references—it gives the strongest combined coverage–grounding profile and improves over its own base on all reported metrics. This indicates stronger grounded keypoint coverage rather than only better reference overlap. Because the supervision comes from a different LLM family and a disjoint document corpus, these gains are not well explained by same-family style matching or memorization of test passages.

Statistical significance. We re-ran every model three times and applied paired bootstrap tests (Berg-Kirkpatrick et al., 2012) over the 1,259 test items (per-item scores averaged across runs; 10,000 resamples, 95% percentile intervals). All six DoRA-SFT-vs-Base contrasts are significant at $p < 0.001$ and survive a Holm correction (Holm, 1979) across the twelve tests in Tables 4 and 23. Two vs. best-baseline contrasts (Table 23) do not: GPT-3.5-Turbo is significantly *better* on BERTScore Recall (-0.55 , $p < 0.05$), and the Irrelevance advantage over GPT-4o ($+1.10$) is not significant ($p=0.066$). Both mark where adaptation stands: the win is in grounded keypoint coverage, not generic embedding similarity.

The gain is not an artifact of answer length. DoRA SFT is the most verbose system in Table 3, so we isolate coverage from length by truncating every answer to $\{0.5, 1, 2, 3\} \times$ its reference length before rescore (Table 26). The advantage over the base survives at every budget, including the $1 \times$ setting where DoRA SFT is by construction *no longer* than the base and still gains $+6.86$ Tok-R.

Metric	$\Delta \uparrow$	95% CI	p
Tok-R	+8.32	[+7.23, +9.42]	<0.001
ROUGE-L-R	+5.74	[+4.65, +6.85]	<0.001
BERTScore-R	+1.10	[+0.78, +1.44]	<0.001
Comp.	+6.50	[+5.19, +7.84]	<0.001
Hall.	+2.60	[+1.91, +3.32]	<0.001
Irrel.	+3.90	[+2.52, +5.28]	<0.001

Table 4: Paired bootstrap tests for DoRA SFT vs. the Llama-3.1-8B base on the 1,259-item benchmark under GTE retrieval (10,000 resamples). Δ is the paired mean *improvement*, so positive means lower Hallucination/Irrelevance; being a per-item paired difference it may deviate by ≤ 0.1 from the difference of rounded means in Table 3. Contrasts against the strongest baseline per metric are in Table 23.

Setting	GPT-4o	Llama-3.1	DoRA SFT	Gain vs Base
GTE	69.40 $\pm$ 0.16	66.72 $\pm$ 0.51	73.37$\pm$0.40	+6.65
IBM Granite	60.50 $\pm$ 0.97	60.69 $\pm$ 0.33	67.20$\pm$0.25	+6.51
Gold evidence	80.36 $\pm$ 0.56	78.25 $\pm$ 0.48	86.25$\pm$0.76	+8.00

Table 5: Downstream QA robustness across retrieval settings, measured by RAGEval Completeness (%). All cells are three-run means $\pm$ std: baselines over three independent inference runs (temperature 0.2, identical frozen context), DoRA SFT over three fine-tuning seeds. Gain vs Base is DoRA SFT minus Llama-3.1.

5.3 Robustness across retrieval settings

The main trend persists across retrieval conditions (Table 5). DoRA SFT tracks retrieval quality—Completeness rises from 67.20% under the weaker IBM Granite retriever to 73.37% under GTE—and peaks under oracle evidence (86.25%), where retrieval error is removed entirely. Adaptation therefore improves not only robustness to imperfect retrieval but also the model’s ability to exploit high-quality evidence (full breakdowns in Appendices A.1 and A.2).

We next condition on *whether evidence was retrieved at all*: Hit@3=1 (809; 64.3%) vs. Hit@3=0 (450; 35.7%). In Table 6, DoRA SFT is strongest when retrieval succeeds (84.38% Completeness, 1.73% hallucination), and its gain over the base *grows* when retrieval misses (+6.17 Completeness, +4.65 hallucination points): the adapted model uses partial or noisy evidence better instead of following irrelevant passages.

5.4 Domain Expert Evaluation

We now report **human evaluation at three stages** as described below. Our human evaluators are defense domain experts who are not authors of the paper and are familiar with the national context of

Model	Hit@3=1			Hit@3=0		
	Comp. $\uparrow$	Hall. $\downarrow$	Irr. $\downarrow$	Comp. $\uparrow$	Hall. $\downarrow$	Irr. $\downarrow$
GPT-4o	81.74	2.81	15.45	46.59	10.63	42.78
Llama-3.1-8B-Inst	79.21	2.96	17.83	46.25	7.11	46.64
DoRA SFT	84.38	1.73	13.89	52.42	2.46	<i>45.12</i>
<i>Gain vs Base</i>	+5.17	+1.23	+3.94	+6.17	+4.65	+1.52

Table 6: Retrieval-conditioned QA under the main GTE setting. Hit@3=1 covers 809 samples (64.3%); Hit@3=0 covers 450 samples (35.7%).

the documents.

Stage 1: Expert validation of synthetically generated QA pairs. Four defense domain experts score a stratified 50-item benchmark sample with the same rubric used by our LLM judge. Inter-annotator Within-1 agreement is 94.7–100% across the seven dimensions and $\kappa_w \geq 0.5$ on three of them (Table 8), so the constructed benchmark items are judged sound by experts who did not build them. Stage 3 reuses these annotations to calibrate the judge itself.

Stage 2: QA Evaluation on Expert-curated QA pairs. We create a separate expert-curated set of 86 samples whose reference answers are written and corrected by defense experts (without any use of LLMs), and are never shown to the evaluated LLMs (Appendix B). At this size the set cannot support precise effect estimates, so we read it as a *directional* external check on data never produced by our pipeline and keep this paper’s quantitative claims anchored on the 1,259-item benchmark. So read, Table 7 points the same way as the synthetic benchmark: DoRA SFT roughly doubles Completeness (24.3% vs. 14.4/14.8%) and roughly halves Hallucination (26.5% vs. 51.7/44.9%), with parallel movement in ROUGE-L and BERTScore Recall. The only weaker signal is Irrelevance: DoRA SFT produces more content outside the gold key-points. Because it also improves Completeness and reduces hallucination, we interpret this as verbosity, not worse factual grounding.

Model	Comp. $\uparrow$	Hall. $\downarrow$	Irr. $\downarrow$	R-L-R $\uparrow$	BERTS-R $\uparrow$
GPT-4o	14.4	51.7	33.9	10.27	65.04
Llama-3.1-8B	14.8	44.9	40.3	18.53	71.43
DoRA SFT	24.3	26.5	49.2	39.20	80.94

Table 7: Expert-curated evaluation. R-L-R=ROUGE-L Recall; BERTS-R=BERTScore Recall.

Dimension	Exact	Within 1	MAD	κ_w
Comprehensiveness	69.3%	97.0%	0.34	0.65
Relevance	82.7%	100.0%	0.17	0.51
Groundedness	85.3%	95.3%	0.19	0.48
Verbosity	65.7%	94.7%	0.40	0.36
Composition	87.0%	99.0%	0.14	0.29
Clarity	87.7%	99.3%	0.13	0.55
Alignment	96.3%	100.0%	0.04	0.16

Table 8: Inter-annotator agreement among four domain-expert annotators on the seven-dimension rubric, computed over the 50 stratified benchmark items.

Stage 3: Expert validation of LLM-as-Judge

To calibrate our LLM-as-judge against expert judgment on the same benchmark items it scores, we conduct a human evaluation on a stratified sample of the DoRA test split. We draw 50 items balanced across the five intent styles (10 each), and four domain-expert annotators score every item independently using seven-dimension quality rubric on a 1–5 ordinal scale (*comprehensiveness*, *relevance*, *groundedness*, *verbosity*, *composition*, *clarity*, *alignment*). Annotators see the (question, evidence, reference answer) triple, score in isolation without seeing the other annotators’ labels, and use a fixed rubric definition document. Table 8 shows strong expert consistency across the seven rubric dimensions: Within-1 agreement is 94.7–100%, mean absolute difference is at most 0.40, and mean pairwise quadratic-weighted κ_w is positive for all dimensions (0.16–0.65). The weakest κ_w is on alignment, where expert scores saturate near the top of the scale, producing a ceiling effect rather than substantive disagreement.

We also calibrate the RAGEval LLM judge against expert scores on the three matching rubric dimensions: *groundedness*, *comprehensiveness*, and *relevance*. On the 50-item sample, the judge aligns best on groundedness ($\kappa_w=0.56$, Within-1 92%), is moderate on comprehensiveness ($\kappa_w=0.23$, Within-1 70%), and is usually within one point on relevance (Within-1 96%, MAD 0.19), where κ_w is suppressed by expert-score saturation (Table 13, Appendix C). These results support the three-layer framing of Section 4: RAGEval grounding signals can be scored by the LLM judge at benchmark scale, while expert review serves as the ground-truth anchor for stylistic and coverage-sensitive dimensions.

5.5 In-Context Learning Baseline

As an adaptation-free baseline, we prompt GPT-4o and Llama-3.1-8B with 1- and 2-shot exemplars curated from DoRA, under the same GTE condition as Table 3. Figure 3 shows only marginal lift from a second exemplar (≤ 2 pp Completeness for both models), leaving DoRA SFT clearly ahead at 73.0%, with parallel gaps on ROUGE-L and BERTScore Recall: curated exemplars supply format guidance, but fine-tuning on the cross-generator split is what buys grounded keypoint coverage.

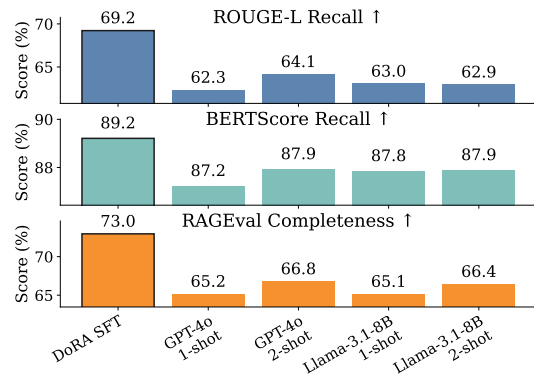


Figure 3: Answer Coverage (ROUGE-L Recall, BERTScore Recall) and RAGEval Completeness for DoRA SFT vs. ICL baselines under GTE retrieval.

5.6 Ablating the Evidence-Selection Pipeline

To identify which stages of the evidence-selection cascade are load-bearing, we rebuild the benchmark four times with everything else fixed (same seed documents, chunker, generator, style mix, quality gate), varying only how E is selected (Appendix J and table 9). *Quality, not yield, separates the methods*: all four configurations yield similar-sized pools, yet randomly chosen bundles are 13.5 points less grounded than the full pipeline as judged by an independent GPT-4o judge. The LLM bundle scorer is the component that matters—removing it costs 3.7–4.4 groundedness points whether the fallback is NLI-only or BM25—while the NLI prefilter alone is the weakest configuration, with the lowest yield and no quality gain over plain BM25. For reuse this means the quality gate cannot compensate for poor evidence selection, and BM25 under the same gate is the reasonable cheap substitute when the LLM scorer is out of budget.

Selection	Pool	Yield $\uparrow$	Grounded $\uparrow$	Answerable $\uparrow$
Full pipeline (ours)	1,237	67.7%	92.5%	93.9%
NLI-only	1,085	56.4%	88.1%	89.2%
BM25 keyword	1,155	68.8%	88.8%	89.6%
Random	1,219	66.3%	79.0%	84.7%

Table 9: Evidence-selection ablation during benchmark construction. Groundedness and answerability are judged by GPT-4o on accepted items only; the quality gate is identical across rows.

6 Discussion

Cross-generator design. The cross-generator + cross-corpus construction (Section 3.4) is DoRA’s most consequential design choice, and transfers to any synthetic-benchmark plus SFT pipeline in a specialist domain at essentially no additional cost—a methodological archetype, not just a defense-domain artifact.

Tight clustering. Under the same retriever, off-the-shelf LMs cluster closely on semantic and grounding diagnostics (Table 3), suggesting that evaluation in private, domain-specific settings is bottlenecked by evidence selection and grounding rather than raw model size. Targeted domain adaptation (even at 8B scale, via LoRA on cross-generator supervision) closes meaningfully more of the visible headroom than swapping in a larger general-purpose model. The benefit is present across every retrieval condition we tested (Table 5): it survives a weaker retriever, and it is largest under oracle evidence, so adaptation is buying better exploitation of whatever evidence is available rather than merely compensating for retrieval noise.

How domain-specific is the taxonomy? Only the seed corpus and the per-style evidence budgets and thresholds are defense-specific. The five intent styles are not: they reduce open-instruction intents to the subset answerable from a document corpus (Section 1), and map onto legal review, clinical guideline lookup, or regulatory compliance unmodified. This is what supports the methodological-archetype claim.

From benchmark to deployment. DoRA is designed for cold-start specialist settings and should be treated as an in-domain regression signal, not a universal leaderboard: teams rerun it after updating the corpus, retriever, prompts, or adaptation, and inspect where failures move across styles and grounding diagnostics. The construction recipe depends

only on partitionable seed documents, an intent taxonomy, and two LLM families. Two caveats bound how far this evidence carries. Every number here comes from one domain: we have not run DoRA on a second specialist corpus, and a credible instantiation in law or biomedicine needs qualified experts in those fields—precisely the resource DoRA assumes is scarce. And the strongest evidence that the framework generalizes beyond its own pipeline, the expert-curated set, rests only on 86 items.

7 Conclusion

We presented DoRA, a domain-grounded benchmark construction and evaluation framework for cold-start RAG in specialist domains. From a small seed set it produces auditable QA instances with linked evidence, enabling evaluation where no in-domain benchmark or labeled dataset exists. Its cross-generator, cross-corpus design—different LLM families over disjoint seed corpora—structurally decouples SFT supervision from the benchmark and reduces the main circularity risks in synthetic evaluation. In a defense case study, LoRA adaptation of Llama-3.1-8B significantly improves grounding diagnostics under the default GTE setting and holds across alternative retrieval conditions, showing that synthetic supervision can improve deployment-facing RAG behavior under domain shift.

Limitations

No cross-domain evaluation. DoRA is presented as a domain-agnostic framework but evaluated on one domain: every number here comes from defense and national-security documents. Transfer to legal, biomedical, or regulated-enterprise corpora is argued from the framework’s structure rather than demonstrated, and a second instantiation would need qualified experts in that field—the scarce resource DoRA is built around. Our results are a case study, not evidence of general transfer.

Residual circularity and generator choice. Different LLM families over disjoint seed documents, with a third model used for judging, close the two most-cited leakage axes, but both splits share one pipeline, so within-pipeline metrics may partly reflect pipeline-level style correlation; the retrieval-conditioned and expert-curated checks mitigate this without ruling it out. We also test one generator

pair (Claude Sonnet 4.6 / GPT-4o, Gemma judging), so portability is a claim about the *design*, not about invariance to the models instantiating it.

Scale and reliability of the evaluation signals. The 86-item expert-curated set and the 50-item, four-annotator calibration suit a setting where expert time binds, but neither supports precise effect estimates; we read the former as directional, not as a leaderboard.

LLM dependence, contamination, and statistical coverage. DoRA relies on capable LLMs for generation, evidence scoring, and evaluation, which can propagate shared biases into the data and its assessment. It measures grounding, not pretraining contamination; since our seed documents are public, adopters in domains with document overlap should add a contamination probe before reading absolute scores. Three-run means and bootstrap intervals cover the main GTE table and the robustness comparison; the retrieval-conditioned split, the full per-metric breakdowns in the appendix, and the expert-curated analysis use one run per baseline and carry no confidence intervals.

Ethics Statement

Sourcing and provenance. The case-study corpus is built entirely from open, publicly released defense material—strategy, capability, procurement, and industry-planning reports—with no classified or proprietary content; what DoRA adapts to is the *language* of the domain, not operational knowledge. Every released instance carries provenance fields, so any QA pair can be audited back to its source passage. Release is at <https://github.com/unswnlporg/DoRA>, itemized in Appendix N under per-component license and sensitivity checks.

Dual use. Defense is our object of study as an under-served evaluation setting, not an operational target: a benchmark, a protocol, and an adaptation recipe are not capabilities specific to military application. The residual concern is adversarial information synthesis: an adapted model could assemble a picture from many public documents faster than an analyst. Three properties bound it: the material is already public and individually retrievable, every answer is evidence-linked and so contestable, and the aggregation is one a diligent reader could perform unaided. We still treat capability-adjacent aggregation as a real if bounded risk, and release a

regeneration recipe rather than pre-assembled corpora where licenses are unclear.

Over-trust and safeguards. The likelier harm is misplaced confidence: a fluent answer invites the reader to skip verification, and the failure mode is real—hallucination is more than halved but not eliminated, and Completeness under retrieval failure stays near 52%. Deployments should surface supporting passages with every answer, flag low confidence when retrieval finds no adequate support, restrict access to vetted users with logging, and never treat an answer as the sole basis for a decision. The system answers questions over a fixed document set; it does not decide, act, or replace human judgment, and human-in-the-loop review is a deployment constraint.

Participation and downstream review. The domain experts who validated the QA pairs, calibrated the judge, and authored the expert-curated set are defense-domain professionals aligned with our industry partners; none are authors, all reviewed only open material, and the released instances carry no personal information. Adopters elsewhere inherit the framework, not this analysis, and should run their own ethics review.

Acknowledgements

This work was funded by Defence Trailblazer through an academia–industry partnership between UNSW and Cyndr. Defence Trailblazer is a part of the Australian Government’s Department of Education Trailblazer Universities Program.

References

- Amirhossein Abaskohi, Spandana Gella, Giuseppe Carenini, and Issam H. Laradji. 2025. [FM2DS: Few-shot multimodal multihop data synthesis with knowledge distillation for question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7256–7282.
- Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, Vignesh P, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, Ken Barker, Martin Franz, Madison Lee, Todd Ward, Salim Roukos, David Cox, Luis Lastras, Jaydeep Sen, and Radu Florian. 2025. [Granite embedding r2 models](#). *Preprint*, arXiv:2508.21085.
- Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. [An empirical investigation of statistical significance in NLP](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language*

- Processing and Computational Natural Language Learning*, pages 995–1005.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17754–17762.
- Mingda Chen, Xilun Chen, and Wen-tau Yih. 2024c. [Few-shot data synthesis for open domain multi-hop question answering](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 190–208, St. Julian’s, Malta. Association for Computational Linguistics.
- André V. Duarte, João DS Marques, Miguel Graça, Miguel Freire, Lei Li, and Arlindo L. Oliveira. 2024. [LumberChunker: Long-form narrative document segmentation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6473–6486, Miami, Florida, USA. Association for Computational Linguistics.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. [Ragas: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating Q&A benchmarks for RAG evaluation in enterprise settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 469–484.
- Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. [Ragbench: Explainable benchmark for retrieval-augmented generation systems](#). *Preprint*, arXiv:2407.11005.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *CoRR*, abs/2312.10997.
- Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>. Accessed 2026-01-22.
- Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if “the moon is made of marshmallows”. *arXiv preprint arXiv:2410.03727*.
- OpenBMB. 2024. MiniCPM-Embedding. <https://huggingface.co/openbmb/MiniCPM-Embedding>. Hugging Face model card; accessed January 23, 2026.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. [KILT: a benchmark for knowledge intensive language tasks](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.
- Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826.

- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. [Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 21999–22027. Curran Associates, Inc.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. [Ares: An automated evaluation framework for retrieval-augmented generation systems](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.
- Maximilian Schmidt, Andrea Bartezzaghi, and Ngoc Thang Vu. 2024. [Prompting-based synthetic data generation for few-shot question answering](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13168–13178, Torino, Italia. ELRA and ICCL.
- Yixuan Tang and Yi Yang. 2024. [Multihop-rag: Benchmarking retrieval-augmented generation for multihop queries](#). *Preprint*, arXiv:2401.15391.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Nandan Thakur, Nils Reimers, Andreas Rückl , Abhishek Srivastava, and Iryna Gurevych. 2021. [Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). *Preprint*, arXiv:2104.08663.
- Prasanth Aby Thomas. 2024. [Meta offers llama ai to us government for national security](#). *CIO*. Accessed 2026-01-22.
- Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Khoshfetrat Pakazad, Tongshuang Wu, and Graham Neubig. 2024. [Synthetic multi-modal question generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12960–12993, Miami, Florida, USA. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.
- Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. [Zerogen: Efficient zero-shot learning via dataset generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Sizhe Yuen, Ting Su, Ziyang Wang, Yali Du, and Adam J. Sobey. 2025. Automatic dataset generation for knowledge intensive question answering tasks. *arXiv preprint arXiv:2505.14212*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. [mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.
- Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. [Rageval: Scenario specific rag evaluation dataset generation framework](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8520–8544, Vienna, Austria. Association for Computational Linguistics.

Appendix

This appendix provides additional results and implementation details for reproducibility:

- **Additional QA Results** (Appendix A): QA with oracle-context (gold evidence) (Appendix A.1 and table 10), and end-to-end QA with IBM Granite retrieval (Appendix A.2 and tables 11 and 12).
- **Manually Curated Dataset** (Appendix B): expert-answered set creation and its use as an external evaluation check.

- **Human-LLM-Judge Alignment on RAGEval-Mapped Dimensions** (Appendix C and table 13): per-dimension calibration of the LLM judge against the mean expert score on the rubric dimensions that map to the RAGEval signals used in the main results.
- **Metrics** (Appendix D): retrieval metrics (Hit@k, Recall@k), retriever identifiers (Table 14), and implementation identifiers for learned metrics and judges (Table 19).
- **RAGEval Metrics** (Appendix E): definitions and formulas for the RAGEval diagnostics (completeness, hallucination, irrelevance) used in our end-to-end QA evaluation.
- **DoRA Dataset** (Appendix F): dataset composition and generation statistics by style.
- **Style Examples** (Appendix G): example questions and answers for each of the five intent styles.
- **Detailed Methodology** (Appendix H): model choices (Table 19) and evidence-selection budgets (Table 20).
- **Construction-Time Quality Control** (Appendix I): composite score weights (Table 21), acceptance thresholds and hard checks (Table 22), and the over-generation yield rate.
- **Evidence-Selection Ablation** (Appendix J and table 9): benchmark quality under four evidence-selection methods with all other pipeline stages held fixed.
- **Full Significance Testing** (Appendix K and table 23): all twelve paired bootstrap contrasts, including those against the strongest baseline per metric.
- **Per-Style Breakdown** (Appendix L and tables 24 and 25): coverage and grounding results disaggregated by intent style.
- **Length-Controlled Coverage Analysis** (Appendix M and table 26): recall metrics recomputed under matched answer-length budgets.
- **Artifact Release and Licensing** (Appendix N and table 27): per-component release plan and licenses.

- **Prompts** (Appendix O): prompt templates used for benchmark-time grounded inference and style-conditioned QA generation.

A Additional QA Results

A.1 QA with Gold Evidence Context

To isolate generation quality from retrieval errors, we also report *gold evidence context* results where the generator is given the recorded gold evidence passages E (oracle context) instead of retrieved passages. Table 10 reports oracle-context results with the same prompt used across models. Under oracle evidence, grounding improves substantially across LLM models; DoRA SFT gives the strongest lexical recall and lowest hallucination rate, while remaining competitive on Completeness and Irrelevance. This indicates that the adaptation benefit is not only a retrieval artifact.

A.2 End-to-End QA with IBM Granite Retriever

We also report end-to-end QA under an alternative fixed retriever, IBM Granite (english-r2), to stress-test robustness to retrieval variation.

In this setting, where IBM Granite is weaker than GTE on retrieval metrics (Figure 2), DoRA SFT *widens its margin on grounding signals*: it achieves the lowest hallucination (3.44%) and second-lowest irrelevance (29.36%) in the aggregate table, more than halving hallucination relative to the base Llama-3.1-8B-Instruct (7.21%). The retrieval-conditioned split in Table 12 shows where the gain comes from. Under successful Granite retrieval, DoRA SFT has the highest Completeness (84.02%) and lowest hallucination (1.95%), improving over the base by +4.09 Completeness points and reducing hallucination by 2.87 points. Under retrieval failure, the margin grows: DoRA SFT reaches 45.74% Completeness, compared with 35.46% for the base and 32.02% for GPT-4o, and reduces hallucination to 4.76%, compared with 10.21% and 15.59%, respectively. This supports the interpretation that DoRA-based adaptation improves retrieval robustness, suitable for real-world applications where retrieval potentially induces errors.

B Manually Curated Dataset

We created a high-quality QA dataset for expert-curated evaluation using a combination of LLM-driven generation, LLM judgment, and human ex-

Model	Len. Ratio	Answer Coverage (Recall)			RAG Faithfulness (LLM judge)		
		ROUGE-L-R $\uparrow$	Tok-R $\uparrow$	BERTScore-R $\uparrow$	Comp. $\uparrow$	Hall. $\downarrow$	Irrel. $\downarrow$
GPT-4o-mini	2.43	75.31	74.35	92.03	83.20	3.40	13.40
GPT-3.5-Turbo	1.61	71.88	71.79	92.11	78.99	2.69	18.32
Llama-3.1-8B-Instruct (base)	1.98	68.22	65.32	89.81	78.23	4.67	17.10
Llama-3-8B-Instruct	2.08	69.47	66.80	90.03	77.52	4.20	18.28
GPT-4o	1.66	72.41	71.51	91.31	79.57	4.23	16.21
Minstral-8B-Instruct	3.84	70.99	59.34	88.59	83.69	3.93	12.38
Claude-3.5-Haiku	3.45	76.23	74.09	90.91	85.64	3.93	10.42
Qwen2.5-7B-Instruct	2.27	74.25	72.87	92.05	81.18	3.53	15.29
DoRA SFT mean$\pm$std	4.17 $\pm$ 0.08	77.61$\pm$0.38	78.28$\pm$0.40	91.68 $\pm$ 0.03	86.25$\pm$0.76	1.96$\pm$0.04	11.79 $\pm$ 0.76
<i>Gain vs Base</i>		+9.39	+12.96	+1.87	+8.02	+2.71	+5.31

Table 10: End-to-end QA with oracle evidence (1,259 test samples). Len. Ratio is generated/gold-answer word-count ratio; ROUGE-L, Tok-R, and BERTScore are recall variants; Comp./Hall./Irrel. are RAGEval diagnostics. All metrics are % except Len. Ratio. For DoRA SFT, we report mean $\pm$ std over three seeds. Gain vs Base is computed from the DoRA SFT mean against Llama-3.1-8B-Instruct (base).

Model	Len. Ratio	Answer Coverage (Recall)			RAG Faithfulness (LLM judge)		
		ROUGE-L-R $\uparrow$	Tok-R $\uparrow$	BERTScore-R $\uparrow$	Comp. $\uparrow$	Hall. $\downarrow$	Irrel. $\downarrow$
GPT-4o-mini	3.16	62.15	62.35	87.22	61.64	7.85	30.51
GPT-3.5-Turbo	1.70	55.49	56.29	86.58	57.36	8.20	34.44
Llama-3.1-8B-Instruct (base)	3.17	62.73	62.07	87.55	60.25	7.21	32.54
Llama-3-8B-Instruct	3.07	61.25	60.63	87.26	59.34	6.57	34.08
GPT-4o	2.04	56.53	56.28	85.32	59.13	9.27	31.60
Minstral-8B-Instruct	4.96	63.15	56.25	86.06	62.58	5.66	31.76
Claude-3.5-Haiku	4.51	65.00	65.59	87.64	65.18	6.94	27.88
Qwen2.5-7B-Instruct	3.34	63.01	62.98	87.94	62.20	6.01	31.79
DoRA SFT mean$\pm$std	5.11 $\pm$ 0.07	66.47$\pm$0.11	68.58$\pm$0.03	88.40$\pm$0.05	67.20$\pm$0.25	3.44$\pm$0.17	29.36 $\pm$ 0.35
<i>Gain vs Base</i>		+3.74	+6.51	+0.85	+6.95	+3.77	+3.18

Table 11: End-to-end QA with IBM Granite retrieval (1,259 test samples, top- $k=3$). Len. Ratio is generated/gold-answer word-count ratio; ROUGE-L, Tok-R, and BERTScore are recall variants; Comp./Hall./Irrel. are RAGEval diagnostics. All metrics are % except Len. Ratio. For DoRA SFT, we report mean $\pm$ std over three seeds. Gain vs Base is computed from the DoRA SFT mean against Llama-3.1-8B-Instruct (base).

Model	Hit@3=1			Hit@3=0		
	Comp. $\uparrow$	Hall. $\downarrow$	Irrel. $\downarrow$	Comp. $\uparrow$	Hall. $\downarrow$	Irrel. $\downarrow$
GPT-4o	80.64	4.26	15.11	32.02	15.59	52.39
Llama-3.1-8B-Inst	79.93	4.82	15.25	35.46	10.21	54.33
DoRA SFT	84.02	1.95	14.02	45.74	4.76	49.51
<i>Gain vs Base</i>	+4.09	+2.87	+1.23	+10.28	+5.45	+4.82

Table 12: Retrieval-conditioned QA under IBM Granite. Hit@3=1 covers 702 samples (55.8%); Hit@3=0 covers 557 samples (44.2%). Comp./Hall./Irrel. are RAGEval diagnostics (%). Gain vs Base is DoRA SFT improvement over Llama-3.1-8B-Inst.

pert annotation. This is a separate, smaller pipeline used only for the expert-curated set summarized in the main paper. We followed a five-step process that blends human expert feedback with LLM-generated data.

Bootstrap dataset. We asked a domain expert to generate 25 QA pairs based on, and answerable using, information from our set of domain-specific seed documents. We obtained 5 examples for each question style: FIND, PROVIDE, GENERATE, SUMMARIZE, and EXPLAIN.

Document chunking. We chunked the seed documents into semantically coherent passages using LumberChunker (Duarte et al., 2024). We used the embedding model listed in Table 19 to calculate an embedding vector for each passage. This process produced 2,187 non-overlapping passages from all seed documents.

Semantic graph construction. We constructed a k -NN graph (for $k = 5$) with nodes corresponding to text passages and edges between two nodes representing semantic similarity estimated using cosine distance between the passages’ embedding vectors. We only accepted edges between nodes when the cosine distance was less than 0.35. The resulting directed graph had 2,187 nodes and 8,121 edges with the average node degree of 3.7 ± 0.9 .

LLM-based dataset generation. We prompted the expert-curated-set generator listed in Table 19 to generate QA pairs for each question style using context retrieved by querying the k -NN graph. First, we selected uniformly at random a node from the graph. Then, depending on the question style, we extracted the ego-graph around the selected

node. For FIND and PROVIDE style questions, we returned only the selected node, and for the other styles we returned the one-hop neighborhood around the node, including the selected node itself. We then prompted the LLM to generate QA pairs for the given question type and context. Figure 4 shows the prompt used for this step. The prompt includes two QA examples for in-context learning selected uniformly at random from the bootstrap dataset conditioned on the question style.

The generated question, answer, and context were then evaluated by the expert-curated-set judge listed in Table 19 using the prompt in Figure 5. We asked it to assign a score in the range $[0, 1]$, where 0 is worst and 1 is best. We accepted generated items with scores above 0.75. This process produced 100 QA pairs with context, with 20 examples for each question style.

Human annotation. Finally, domain experts were shown the generated questions and context and asked to provide an answer. We did not show the generated answer to the annotators. Of the 100 generated examples, the annotators deemed 86 valid and provided answers derived from the supplied passages. The combination of LLM-generated questions, human-generated answers, and the relevant context comprises the final dataset.

B.1 Use as an External Evaluation Set

The final 86 expert-answered items are used only as an external evaluation check, not as supervised fine-tuning data. This set complements the larger synthetic benchmark by testing whether the same systems remain grounded when evaluated against independently written expert answers. We evaluate models under both retrieved context and oracle gold evidence with the same semantic and RAGEval-style diagnostics used in the main experiments; results are summarized in Table 7.

C Human–LLM-Judge Alignment on RAGEval-Mapped Dimensions

Our main results report RAGEval Completeness, Hallucination, and Irrelevance, scored by an LLM judge following the open-source RAGEval protocol of Zhu et al. (2025). To calibrate this judge against expert judgment we focus on the three rubric dimensions that correspond directly to those signals—*groundedness* (hallucination counterpart), *comprehensiveness* (completeness counterpart), and *relevance* (irrelevance counterpart)—and score the

Dim.	Exact	Within-1	MAD	κ_w
Comp.	24.0%	70.0%	0.94	0.23
Hall.	80.0%	92.0%	0.30	0.56
Irrel.	80.0%	96.0%	0.19	−0.07

Table 13: Human–LLM-judge alignment on the 50-item stratified sample. Exact, Within-1, and MAD compare the mean expert score with the LLM judge; κ_w is quadratic-weighted Cohen’s kappa.

same 50-item stratified sample described in Section 5.4 under the identical rubric. Table 13 reports the alignment between the LLM judge and the mean score across the four domain-expert annotators on these three RAGEval-mapped dimensions. The remaining four rubric dimensions (verbosity, composition, clarity, alignment) are stylistic and have no RAGEval counterpart; they are omitted from this calibration. The judge agrees substantially with the experts on *groundedness* ($\kappa_w=0.56$, Within-1 92%), moderately on *comprehensiveness* ($\kappa_w=0.23$, Within-1 70%), and very tightly within one point on *relevance* (Within-1 96%, MAD 0.19), where κ_w is near zero because the expert scores saturate at 4–5 and leave little variance for chance correction. As noted in the main paper, we therefore treat LLM-judge metrics as scalable diagnostics for grounding-anchored signals rather than as a substitute for expert review on borderline items.

D Metrics

For each benchmark instance i , we treat the question q_i as the retrieval query, and the gold evidence passages recorded in DoRA as the set of relevant items G_i (one or more chunk IDs). A retriever produces a ranked list of retrieved passages $R_k(q_i) = (d_{i,1}, d_{i,2}, \dots, d_{i,k})$, where $d_{i,j}$ is the chunk ID at rank j for query q_i , and k is the evaluation cutoff (e.g., $k = 3$ in our QA study).

Hit@ k . Let I be the set of evaluated (e.g. DoRA test set). We define the instance-level indicator

$$\text{Hit@}k_i = \mathbf{1}[\exists j \in \{1, \dots, k\} : d_{i,j} \in G_i],$$

and report the average $\text{Hit@}k = \frac{1}{|I|} \sum_{i \in I} \text{Hit@}k_i$. Intuitively, Hit@ k measures whether the retriever returns *any* gold evidence chunk in its top- k list for q_i .

Recall@ k . When multiple evidence passages are marked relevant, we measure coverage of the gold

You are a helpful assistant tasked with creating a single-turn question answering dataset. Your task is to generate one question-answer (QA) pair based on the provided text passages.

****Instructions:****

- The question must be answerable using only the information in the passages.
- The question type must match the specified type below.
- Use the passage references ([1], [2], etc.) to ground your question and answer.
- Do not use external knowledge or make assumptions beyond the passages.
- Ensure the answer is concise, factually correct, and directly supported by the text.

****Question Type:**** {question_type}

****Examples:****

{examples}

****Passages:****

{passages}

****Output Format (JSON):****

```
{
  "question": "Your generated question here",
  "answer": "Your generated answer here"
}
```

Figure 4: Prompt template used for generating QA pairs with In-Context Learning.

set:

$$\text{Recall@}k_i = \frac{|\{d_{i,1}, \dots, d_{i,k}\} \cap G_i|}{|G_i|},$$

and report $\text{Recall@}k = \frac{1}{|I|} \sum_{i \in I} \text{Recall@}k_i$. This measures how completely the top- k retrieved chunks cover the gold evidence set (with set intersection preventing repeated retrieval of the same chunk from inflating the score).

Details of retrievers are shown in Table 14.

Answer-coverage metric models. For reproducibility of the learned answer-coverage scores, the implementation identifier for BERTScore Recall is listed in Table 19.

E RAGEval Metrics

We compute the end-to-end faithfulness diagnostics using the open-source RAGEval evaluation implementation³ from Zhu et al. (2025), with the rubric-judge setup listed in Table 19. For each instance i , RAGEval extracts a set of reference keypoints $\mathcal{K}_i = \{k_{i,1}, \dots, k_{i,N_i}\}$ from the gold answer a_i and evaluates a system answer $\hat{a}_i$ by classifying each keypoint into one of three categories:

- $\mathcal{K}_{i,\text{comp}}$: keypoints judged as *complete* (covered correctly by $\hat{a}_i$);

- $\mathcal{K}_{i,\text{irr}}$: keypoints judged as *irrelevant* (missing/unaddressed by $\hat{a}_i$); and
- $\mathcal{K}_{i,\text{hall}}$: keypoints judged as *hallucinated/wrong* (addressed incorrectly by $\hat{a}_i$).

Instance-level scores. The metrics are defined as normalized keypoint counts:

$$\begin{aligned} \text{RAG-Comp}_i &= \frac{|\mathcal{K}_{i,\text{comp}}|}{|\mathcal{K}_i|}, \\ \text{RAG-Hall}_i &= \frac{|\mathcal{K}_{i,\text{hall}}|}{|\mathcal{K}_i|}, \\ \text{RAG-Irr}_i &= \frac{|\mathcal{K}_{i,\text{irr}}|}{|\mathcal{K}_i|}. \end{aligned}$$

All three scores lie in $[0, 1]$; higher completeness is better, while lower hallucination and irrelevance are better.

Dataset-level reporting. We report micro-averages over instances:

$$\text{RAG-Metric} = \frac{1}{|I|} \sum_{i \in I} \text{RAG-Metric}_i,$$

where I is the evaluated set (e.g., the 1,259 test split from DoRA dataset).

³<https://github.com/OpenBMB/RAGEval/tree/main/rageval/evaluation>

You are an expert evaluator tasked with assessing the quality of a question-answer (QA) pair generated from a set of passages. Your evaluation should determine how well the QA pair aligns with the intended question type and whether it is accurate and contextually grounded.

Please assess the QA pair based on the following criteria:

1. **Question Type Alignment**: Does the question clearly reflect the specified question type (e.g., Find, Explain, Summarize)?
2. **Contextual Answerability**: Can the question be answered using only the information provided in the passages?
3. **Answer Accuracy**: Is the answer factually correct and directly supported by the passages?

Provide a brief reasoning for your evaluation and assign a **quality score** between 0.0 (poor) and 1.0 (excellent).

```

Question Type: {question_type}

Passages:
{passages}

Question: {question}

Answer: {answer}

Output Format (JSON):
{
  "reasoning": "Your reasoning about the evaluation and score",
  "score": 0.85
}

```

Figure 5: Prompt template used for judging the quality of generated question-answer pairs conditioned on the question style.

Retriever	Model / Implementation ID
BM25 (Robertson and Zaragoza, 2009)	rank-bm25
GTE-multilingual-Base (Zhang et al., 2024)	Alibaba-NLP/gte-multilingual-base
MiniCPM-Embedding (OpenBMB, 2024)	openbmb/MiniCPM-Embedding
BGE-M3 (Chen et al., 2024a)	BAAI/bge-m3
IBM Granite (english-r2) (Awasthy et al., 2025)	ibm-granite/granite-embedding-english-r2

Table 14: Retriever identifiers for reproducibility.

F DoRA Dataset

Details regarding the train/test sets used to benchmark different LLMs in the defense domain are shown in Tables 16 and 17.

G Example Questions and Answers by Style

We use five intent styles to approximate deployment traffic in defense settings:

- **FIND**: pinpointing exact data points (e.g., deployment dates, cost figures);
- **EXPLAIN**: clarifying technical distinctions critical to planning and risk assessment;
- **SUMMARIZE**: condensing lengthy reports into concise, actionable overviews;

- **GENERATE**: producing structured lists of capabilities, gaps, or priorities at a glance; and
- **PROVIDE**: surfacing key quantitative metrics (budgets, procurement data).

Examples for each style are shown in Table 18.

H Detailed Methodology

Core model choices. Table 19 summarizes the default model choices by pipeline stage.

Computational budget and infrastructure. Most benchmark construction and evaluation stages use API-hosted LLMs or lightweight local encoders, so exact provider-side compute is not observable. For local components, we use the model identifiers listed in Table 19; the largest local model

Work	S	E	MH	MM	Scn	Sty	Seed	Notes
Few-shot Data Synthesis for Open-Domain Multi-hop QA (EACL'24) (Chen et al., 2024c)	✓	✗	✓	✗	Open	✗	✓	Few-shot pipeline synthesizing multi-hop QA from Wikipedia document pairs; improves smaller LMs on Hotpot-style tasks; no dedicated RAG evaluation metrics.
FM2DS (EMNLP'25) (Abaskohi et al., 2025)	✓	✗	✓	✓	Open	✗	✓	Few-shot multimodal, multihop synthesis grounded in multimodal Wikipedia pages; introduces a multimodal multi-hop benchmark over long docs.
SMMQG (Findings EMNLP'24) (Wu et al., 2024)	✓	✗	✗	✓	Open	✓	✓	Generates modality/style-conditioned multimodal questions grounded in Wikipedia documents to stress MMRAG; not explicitly multi-hop.
Prompting-based SDG (LREC-COLING'24) (Schmidt et al., 2024)	✓	✗	✗	✗	Open	✗	✗	Prompting-based synthetic QA generation for few-shot settings; not grounded to specific seed docs.
RAGAS (EACL'24) (Es et al., 2024)	✗	✓	✗	✗	Open	N/A	N/A	Reference-free RAG metrics (faithfulness, answer/context relevance); framework only.
ARES (NAACL'24) (Saad-Falcon et al., 2024)	✗	✓	✗	✗	Open	N/A	N/A	Automated evaluation framework with faithfulness/attribution metrics and diagnostics.
RAGChecker (NeurIPS'24 D&B) (Ru et al., 2024)	✗	✓	✗	✗	Open	N/A	N/A	Fine-grained diagnostic framework for RAG (answerability, support, hallucination).
RAGEval (ACL'25) (Zhu et al., 2025)	✓	✓	✗	✗	Scn.	✗	✗	Scenario templates derived from seed docs, but QA/eval built over generated articles/configs rather than directly over the original seed docs.
Automatic Dataset Generation for KIQA (arXiv'25) (Yuen et al., 2025)	✓	✗	✗	✗	Domain	✗	✓	Generates domain QA from technical sources (e.g., manuals/specs). Emphasis on creating fine-tuning data rather than defining an evaluation framework.
DataMorgana (ACL'25 Ind.) (Filice et al., 2025)	✓	✗	✗	✗	Domain	✗	✓	Enterprise benchmark synthesis over private corpora with user-defined question categories to control traffic; emphasizes diversity and question–passage fit, leaving evidence-grounded answer scoring largely out of scope.
Ours (DoRA)	✓	✓	✗ [†]	✗ [†]	Domain	✓	✓	Domain-oriented, contamination-aware evaluation with style-conditioned synthetic Q&A grounded in domain sources; task-aware scoring and baselines.

Table 15: Comparison across synthetic data generation, RAG evaluation, realistic context dimensions, and whether question generation is grounded in seed documents. Abbrev.: S=Synthetic, E=RAG Evaluation, MH=Multi-hop, MM=Multimodal, Scn=Scenario, Sty=Style, Seed=Seed-doc. N/A marks evaluation frameworks rather than data-generation methods. [†]Multi-hop, Multimodal are configurable (planned stressors) rather than a core requirement.

Task	Samples	%	Docs
EXPLAIN	1147	21.5%	20
FIND	986	18.5%	20
GENERATE	1184	22.2%	20
PROVIDE	873	16.4%	20
SUMMARIZE	1133	21.3%	20
Total	5,323	100.0%	20

Table 16: Cross-generator training-set composition by style (5,323-instance Claude Sonnet 4.6 training split, 90/10 train/eval), used for DoRA SFT.

used during construction is Gemma-2-9B-IT for evidence scoring, run in 8-bit mode, while the main adapted generator is Llama-3.1-8B-Instruct with LoRA adapters. DoRA SFT is performed on a single NVIDIA GPU workstation; across the three fine-tuning seeds reported in Table 3, the total training budget is approximately 10 GPU-hours of 1xA100 GPU (80GB). We report API model names rather than parameter counts where providers do not disclose exact deployed model sizes, and we release prompts, code, and configuration files to make the observable compute path reproducible.

Training setup and hyperparameters. For DoRA SFT, we fine-tune Llama-3.1-8B-Instruct with LoRA adapters on the 5,323-instance cross-generator training split, using the 90/10 train/eval partition described in Table 16. We use the same retrieved-context input format as the downstream QA experiments: the question is concatenated with the evidence context, and the model is trained to generate the reference answer. We tune only a small set of LoRA and optimization hyperparameters on the held-out eval split, selecting the checkpoint with the best validation loss before evaluating once on the fixed test split. The best-found configuration uses LoRA rank 64, LoRA $\alpha=128$, dropout 0.05, learning rate 1×10^{-5} , batch size 4 (effective 16 with gradient accumulation), gradient accumulation 4, max sequence length 4096, and 3 training epochs (checkpoint selected by best validation loss). The three-seed results in Table 3 rerun this selected configuration with different random seeds rather than re-tuning per seed.

Tokenization and chunking. We tokenize with c1100k_base. Chunks are built by packing paragraphs to a target budget of 512 tokens (min 100;

Task	Samples	%	Avg Q Len	Avg A Len	Refs/Sample	Duration (s)	Docs
EXPLAIN	165	13.1%	13.67	33.59	2.99	14.06	18
FIND	421	33.4%	12.65	20.83	1.00	4.79	20
GENERATE	309	24.5%	13.76	31.42	3.34	14.31	20
PROVIDE	155	12.3%	14.00	30.45	1.91	12.37	19
SUMMARIZE	209	16.6%	12.85	29.35	2.65	11.29	18
Total/Avg	1259	100.0%	13.26	27.70	2.22	10.35	~19

Table 17: Test set composition and basic statistics by style. “Refs/Sample” counts linked evidence passages.

Style	Example question	Example answer
FIND	What capabilities does the Research Technology and Operations provide to Defense IS&T?	Research Technology and Operations provides specialist technical and business capabilities to Defense IS&T.
EXPLAIN	How is the concept of ‘asymmetric advantage’ defined in a military context?	In a military context, ‘asymmetric advantage’ refers to military capabilities that pit strength against weakness, often in a non-traditional or unconventional manner, and that disrupt a potential adversary’s decision calculus.
SUMMARIZE	What initiatives are being continued and enhanced to support the growth of the defense industry workforce?	The Schools Pathways Program, Defense Industry Internship Program, and Defense Industry Pathways Program will continue and be enhanced to achieve better outcomes.
GENERATE	What measures will be taken to provide the industry with clearer information about Defense’s needs?	The industry will receive clearer and more detailed information about Defense’s needs through more classified briefings, the establishment of a tri-partite defense industry council, and industry forums.
PROVIDE	What was the total value of contracts awarded by Defense in 2022–23?	In 2022–23, Defense awarded over \$38 billion in contracts.

Table 18: Illustrative examples of the five DoRA intent styles. Examples are shown for readability and to convey the expected output format.

max 1024). Oversized chunks trigger a sentence-level sliding-window fallback with $\approx 25\%$ sentence overlap. Each chunk retains provenance (document id/name/path, page numbers, optional section header).

Retrieval. During dataset construction, candidate passage retrieval uses a hybrid dense+BM25 stage with cosine similarity (L2-normalized embeddings, inner-product search) and score mixing (0.7 dense / 0.3 BM25) before NLI-based evidence filtering. For retrieval-conditioned benchmark runs, we instead use the fixed retrievers described in the main paper (e.g., GTE-multilingual-Base or IBM Granite), with chunk size 320 tokens, overlap 64, and top- $k=3$.

Evidence selection and budgets. Evidence selection is conditioned on style s . We first retrieve candidates, then apply an NLI prefilter of size K , and finally use a local LLM scorer to rank candidate *combinations* for completeness, complementarity, coherence, and task fit. Table 20 summarizes the per-style evidence budgets and selection hyperparameters. When multi-chunk synthesis is enabled, we require at least two distinct evidence chunks for long-form styles.

Style	$ E $	K	Comb.	Notes
FIND	1	20	6	Extractive
PROVIDE	2	30	7	Numeric content required
EXPLAIN	4	60	8	Long-form reasoning
SUMMARIZE	4	40	8	Coverage-oriented
GENERATE	5	50	7	List/category synthesis

Table 20: Evidence-selection budgets by style. $|E|$ is the target evidence-bundle size; K is the NLI prefilter size; Comb. is the maximum number of chunk combinations evaluated.

Construction-time filtering. We over-generate ($3\times$ target size), remove exact duplicate questions per seed document, and apply task-aware quality control. Hard gates enforce non-empty fields and numeric content for PROVIDE; for multi-chunk items, accepted candidates must draw from at least two distinct evidence chunks. Evidence-grounded scoring uses NLI, span alignment, numeric consistency, and semantic relevance/coverage with the models listed above. We then stratify candidates by style and select the highest-quality items under per-style quotas.

I Construction-Time Quality Control: Weights, Gates, and Yield

This section reports the concrete values behind the acceptance rule of Section 3, so that a practitioner can calibrate their own DoRA instantiation rather

Stage	Model / setting
Q&A generation – test split	openai/gpt-4o (temp 0.3, max 512 new tokens)
Q&A generation – training split	anthropic/claude-sonnet-4.6 (temp 0.3, max 512 new tokens)
Expert-curated set: Q&A generation	openai/gpt-4o
Expert-curated set: quality judge	gemini-2.5-pro
Expert-curated set: graph encoder	sentence-transformers/all-MiniLM-L6-v2
Human–LLM-judge calibration	openai/gpt-4o-mini via OpenRouter (temp 0)
RAGEval rubric judge	openai/gpt-4o-mini
Retrieval encoder (construction)	sentence-transformers/all-mpnet-base-v2
Retrieval encoder (retrieval-conditioned runs)	Alibaba-NLP/gte-multilingual-base
Evidence selection (NLI prefilter)	facebook/bart-large-mnli
Evidence selection (local scorer)	google/gemma-2-9b-it (8-bit)
Faithfulness: entailment (DES)	microsoft/deberta-large-mnli
Faithfulness: extractive span QA	deepset/roberta-base-squad2
Faithfulness: embedding similarity	sentence-transformers/all-MiniLM-L6-v2
Answer coverage: BERTScore Recall	distilbert-base-uncased

Table 19: Default model and metric implementation choices by stage under the cross-generator, cross-corpus design. The training and test splits use *different LLM families* over *disjoint* seed-document corpora.

than re-deriving them.

Composite score weights. The composite quality score Q_s is a per-style weighted sum over five faithfulness components: NLI entailment against the evidence bundle (NLI), extractive span alignment (Span-F1), numeric consistency (Num), keypoint recall (Rec), and question–answer relevance (Rel). Table 21 lists the weights $w_{s,m}$. The extractive styles put roughly 80% of the mass on entailment and span alignment, whereas the synthesis styles put roughly 70% on keypoint recall and relevance; PROVIDE is the only style with substantial numeric-consistency weight.

Style s	NLI	Span-F1	Num	Rec	Rel
FIND	0.45	0.35	0.15	0.00	0.05
PROVIDE	0.35	0.25	0.35	0.00	0.05
EXPLAIN	0.20	0.15	0.00	0.25	0.40
SUMMARIZE	0.20	0.10	0.00	0.35	0.35
GENERATE	0.20	0.10	0.00	0.35	0.35

Table 21: Composite quality-score weights $w_{s,m}$ by style. Component models: NLI = microsoft/deberta-large-mnli; extractive QA (Span-F1) = deepset/roberta-base-squad2; embeddings (Rel) = sentence-transformers/all-MiniLM-L6-v2.

Acceptance thresholds. A candidate is accepted iff both conjuncts of the rule in Section 3 hold: the style-specific hard checks $h_s(q, a, E) = 1$ and the composite threshold $Q_s(q, a, E) \geq \tau_s$. In our defense instantiation $\tau_s=0.46$ for all five styles, so per-style strictness is expressed entirely through h_s .

Two hard checks are universal: non-empty question, answer, and evidence fields; and, for multi-chunk styles, accepted candidates must draw from at least two distinct evidence chunks. The style-specific checks are in Table 22.

Style	Grounding floor	Numeric requirement	Min. keypoint recall
FIND	Span-F1 ≥ 0.13	Numeric consistency ≥ 0.80 (when numbers present)	0.75
PROVIDE	Span-F1 ≥ 0.13	Must contain numeric content; consistency ≥ 0.80	0.75
EXPLAIN	NLI ≥ 0.025	—	0.65
SUMMARIZE	NLI ≥ 0.025	—	0.65
GENERATE	NLI ≥ 0.025	—	0.60

Table 22: Style-specific hard checks h_s . All listed conditions must hold, in addition to the two universal gates described in the text.

Yield. We set the over-generation budget to a maximum of $3\times$ the target size. The training campaign produced 9,891 candidate QA pairs, of which 5,323 passed exact-duplicate removal and task-aware quality control—a yield of **53.8%**. The accepted 5,323 instances are the training split reported in Table 16.

J Evidence-Selection Ablation

DoRA’s evidence selection is a three-stage cascade: an NLI prefilter reduces the candidate pool, a local LLM scorer ranks candidate bundles on completeness, complementarity, coherence, and task fitness, and the highest-scoring bundle is returned. To establish which stages are necessary, we rebuild the benchmark four times with everything else held fixed—the same 20 seed documents, chunker, generator (Claude Sonnet 4.6), style mix, and quality gate—varying only the selection method: the

full pipeline; NLI prefiltering with no LLM scorer; BM25 keyword matching; and random bundle selection.

We report *yield*, the share of generated candidates that pass the quality gate, and, on the accepted items only, *groundedness* and *answerability* as scored by an independent judge (GPT-4o), which is used nowhere in generation or selection. Yield measures how much data a configuration produces; groundedness and answerability measure whether what survives the gate is usable.

Table 9 separates the two. Yield is nearly flat across the full pipeline, BM25, and random selection (66–69%), and is *lowest* for NLI-only (56.4%), so no configuration is distinguished by how much data it produces. Quality is where they diverge: the full pipeline is best on both groundedness (92.5%, +3.7 over the best simplified method) and answerability (93.9%, +4.3), and random selection—which still passes two thirds of its candidates through the gate—produces items that are 13.5 groundedness points worse. That gap is the clearest statement of the result: a quality gate applied downstream cannot repair evidence that was poorly chosen upstream.

Among the simplifications, the ordering is informative. Dropping only the LLM scorer (NLI-only, BM25) costs 3.7–4.4 groundedness points, whereas dropping selection entirely (random) costs 13.5, so the bundle-scoring step carries most of the benefit and the prefilter contributes mainly by narrowing the search. NLI-only is not a viable simplification: it has the lowest yield of the four and no quality advantage over plain BM25 keyword matching. For a practitioner under compute constraints, BM25 with the same quality gate is therefore the sensible cheap substitute, at a cost of roughly four groundedness points relative to the full cascade.

K Full Significance Testing

Table 4 in the main paper reports the six DoRA-SFT-vs-Base contrasts. Table 23 adds, for each metric, the contrast against the strongest baseline on that metric. All twelve tests use the same protocol: per-item scores averaged over each model’s three runs, paired bootstrap resampling (Berg-Kirkpatrick et al., 2012) with 10,000 replicates over the 1,259 test items, and 95% percentile intervals. Ten of the twelve contrasts remain significant under a Holm correction (Holm, 1979); the two exceptions are reported here rather than omitted, since

they identify precisely where adaptation confers no advantage.

Metric	Contrast	$\Delta \uparrow$	95% CI	<i>p</i>
Tok-R	vs Base	+8.32	[+7.23, +9.42]	<0.001
Tok-R	vs GPT-4o-mini	+3.74	[+2.76, +4.73]	<0.001
ROUGE-L-R	vs Base	+5.74	[+4.65, +6.85]	<0.001
ROUGE-L-R	vs GPT-4o-mini	+2.08	[+1.00, +3.15]	<0.001
BERTScore-R	vs Base	+1.10	[+0.78, +1.44]	<0.001
BERTScore-R	vs GPT-3.5-Turbo	−0.55	[−0.87, −0.21]	<0.05
Comp.	vs Base	+6.50	[+5.19, +7.84]	<0.001
Comp.	vs GPT-4o	+3.93	[+2.54, +5.31]	<0.001
Hall.	vs Base	+2.60	[+1.91, +3.32]	<0.001
Hall.	vs GPT-3.5-Turbo	+1.48	[+0.81, +2.20]	<0.001
Irrel.	vs Base	+3.90	[+2.52, +5.28]	<0.001
Irrel.	vs GPT-4o	+1.10	[−0.34, +2.55]	0.066

Table 23: All twelve paired bootstrap contrasts for DoRA SFT under GTE retrieval. Δ is the paired mean *improvement*, so a positive value means lower Hallucination/Irrelevance. Base is Llama-3.1-8B-Instruct.

L Per-Style Breakdown

The main tables aggregate over the five intent styles, and FIND is the largest stratum of the test split (421 of 1,259 items; Table 17), so aggregate gains could in principle be driven by a single style. Tables 24 and 25 show they are not. DoRA SFT improves over the base on every style on every metric, and matches or exceeds GPT-4o on completeness and hallucination for all five—including the synthesis-heavy GENERATE and SUMMARIZE styles, where an adapted 8B model has the least structural advantage. The gain is largest on FIND (+13.28 Token Recall), which is consistent with extractive questions having a single well-defined gold span that adaptation can learn to target, and smallest but still substantial (+5.33 to +5.55) on the long-form styles.

Style	Base	GPT-4o	DoRA SFT	Gain vs Base
FIND	63.24±1.21	69.68±0.15	76.52±0.26	+13.28
PROVIDE	58.46±1.15	61.57±0.20	66.39±0.27	+7.93
EXPLAIN	63.46±0.34	67.93±0.25	68.79±0.54	+5.33
SUMMARIZE	64.85±0.98	64.68±0.63	70.19±0.25	+5.34
GENERATE	62.50±0.38	63.53±0.20	68.05±0.30	+5.55

Table 24: Token Recall ($\uparrow$) by intent style under GTE retrieval; mean±std over three runs per model. Base is Llama-3.1-8B-Instruct.

M Length-Controlled Coverage Analysis

DoRA SFT produces longer answers than every baseline (Length Ratio 5.09; Table 3), which raises the question of whether its recall-oriented gains are a mechanical consequence of generating more

Style	Completeness $\uparrow$			Hallucination $\downarrow$		
	Base	GPT-4o	DoRA SFT	Base	GPT-4o	DoRA SFT
FIND	73.58 $\pm$ 0.92	74.23 $\pm$ 0.16	80.25$\pm$0.11	3.75 $\pm$ 0.41	3.80 $\pm$ 0.50	1.55$\pm$0.25
PROVIDE	58.33 $\pm$ 0.62	63.11 $\pm$ 0.64	66.16$\pm$1.07	7.64 $\pm$ 0.53	6.84 $\pm$ 1.21	3.81$\pm$0.74
EXPLAIN	62.83 $\pm$ 1.34	69.02 $\pm$ 0.77	71.11$\pm$0.84	5.86 $\pm$ 0.49	4.55 $\pm$ 0.17	2.76$\pm$0.57
SUMMARIZE	67.00 $\pm$ 1.54	68.85 $\pm$ 0.83	72.23$\pm$0.35	3.55 $\pm$ 0.15	5.45 $\pm$ 0.82	2.26$\pm$0.52
GENERATE	63.57 $\pm$ 0.54	66.62 $\pm$ 0.14	69.67$\pm$0.60	4.85 $\pm$ 0.59	5.38 $\pm$ 0.45	1.84$\pm$0.36

Table 25: RAGEval Completeness and Hallucination by intent style under GTE retrieval; mean $\pm$ std over three runs per model.

tokens. Recall metrics do not penalize extra tokens, so more output cannot lower the score—but it can raise it by chance overlap. To separate coverage from length we truncate every generated answer to a per-sample budget of $\{0.5, 1, 2, 3\} \times$ its reference length before recomputing the metrics, using the same implementations as the main tables.

Table 26 shows the advantage over the base survives at every budget. The decisive row is $1\times$: answers are capped at the gold answer’s length, so DoRA SFT is by construction *no longer* than the base, and it still gains +6.86 Token Recall. ROUGE-L Recall shows the same pattern, with gains over the base of +4.7/+6.4/+3.1/+4.3/+5.9 across the five budgets. Length therefore cannot be the explanation for the improvement; this is consistent with the design argument in Section 4—the RAGEval judge scores discrete gold keypoints, so unsupported padding registers as Irrelevance, and Hallucination falls by more than half rather than rising as token accumulation would predict.

Budget	Base	GPT-4o	DoRA SFT	Gain vs Base
0.5 $\times$	25.07 $\pm$ 0.07	29.20$\pm$0.07	29.07 $\pm$ 0.37	+4.00
1 $\times$	43.26 $\pm$ 0.37	49.45 $\pm$ 0.04	50.12$\pm$0.62	+6.86
2 $\times$	57.40 $\pm$ 0.71	62.43 $\pm$ 0.21	63.07$\pm$0.41	+5.67
3 $\times$	60.22 $\pm$ 0.78	64.62 $\pm$ 0.16	67.21$\pm$0.34	+6.99
Untruncated	62.76 $\pm$ 0.74	66.09 $\pm$ 0.15	71.13$\pm$0.17	+8.37

Table 26: Token Recall ($\uparrow$) under matched answer-length budgets. Each generated answer is truncated to the stated multiple of its own reference answer’s length before scoring; mean $\pm$ std over three runs per model.

N Artifact Release and Licensing

All artifacts are released at <https://github.com/unswnlporg/DoRA>. The 40 seed documents are re-distributed directly as a PDF bundle under their publishers’ open-access terms. For the derived instances we distinguish between components that are *directly redistributable* and the residual ones released as a *regeneration recipe*: any instance whose derivation would redistribute source text under a license that does not permit it is released as

a pointer triple (seed-document identifier, prompt, generation config) plus the recorded provenance fields, so an external team can regenerate the identical item locally. Everything needed to rebuild the full benchmark end-to-end is in one of the two categories—no component is withheld.

Component	Form	License	Notes
Seed documents (40 PDFs)	Direct	Original publisher terms (open public release)	Redistributed as a bundle under the publishers' open-access terms, with the source list retained for provenance
Generation, evidence-scoring, and quality-control prompts and templates	Direct	CC BY 4.0	Complete set, matching Appendix O
Construction, SFT, retrieval, and evaluation code	Direct	Apache 2.0	Implements the three-layer metric stack. The RAGEval keypoint judge is an <i>external dependency</i> , obtained separately by the user (required to reproduce the RAG-faithfulness columns) and licensed CC BY-NC-SA 4.0; reproducing our RAGEval-scored results therefore inherits its Non-Commercial term
Derived benchmark instances (1,259 test; 5,323 train) with provenance fields	Direct where source license permits; otherwise regeneration recipe	CC BY 4.0 (derived fields)	Provenance = question, answer, style, evidence bundle, source passage IDs
Expert-curated evaluation set (86 items)	Direct	CC BY 4.0	Expert-authored answers; no source text redistributed
Human-evaluation rubric and de-identified annotator scores	Direct	CC BY 4.0	Supports Tables 8 and 13
LoRA adapter weights and training configuration	Direct	Llama 3.1 Community License	Derivative of Llama-3.1-8B-Instruct; inherits the base-model license

Table 27: Per-component release plan and licensing. “Regeneration recipe” means a seed-document pointer, the exact prompt, and the generation configuration sufficient to reconstruct the item locally without redistributing licensed source text.

O Prompts

O.1 Multi-Chunk Synthesis Directives

For multi-chunk synthesis, we append the following directives (with configuration-dependent values, e.g., minimum chunks/citations) to each style prompt:

```
MULTI-CHUNK SYNTHESIS MODE (ENFORCED):  
- Integrate evidence from at least 2 distinct [Chunk] sections when they offer relevant details.  
- Cite each supporting chunk explicitly by number or by quoting a short anchor phrase (minimum 2 citations).  
- Merge overlapping insights into one cohesive answer rather than separate mini-answers.  
- Reiterate key facts at the start of the answer so earlier chunks are not ignored.  
- Answers lacking multi-chunk grounding will be rejected by downstream verification.
```

O.2 Benchmark-Time Grounded Inference Prompt

For end-to-end QA evaluation, each model receives the retrieved context (or oracle gold evidence) and the test question using the following grounded inference prompt:

```
You are provided with grounded evidence. Use only this context to answer.  
If the evidence is insufficient, respond with "I don't have enough context."  
  
Context:  
{context}  
  
Question:  
{question}  
  
Answer:
```

O.3 Style-Conditioned Q&A Generation Prompts

We generate each candidate instance by prompting the dataset-construction model listed in Table 19 with an evidence context and a style-specific template. In all prompts below, {context} is replaced with the selected evidence passages (concatenated).

FIND.

Based on the provided context, generate a factual question-answer pair that asks for specific information (dates, numbers, names, locations, etc.) that can be found in the text.

Context:
{context}

Requirements:

- Question should ask for specific factual information
- Answer should be a short, precise factual response directly from the text
- Copy the shortest exact phrase that answers the question (verbatim) from the supporting sentence
- If needed, include a brief quoted clause that contains the fact
- Answer must be declarative and must not be phrased as a question; do not include a question mark
- The answer must exist exactly in the provided context
- Focus on concrete details like dates, amounts, names, locations, or specific facts
- Prefer interrogatives: What/When/Where/Who/Which/How many/How much (avoid standalone 'How')
- If multiple values appear, choose the one that most directly answers the question
- Use only information present in the Context; do not add external knowledge
- Do not refer to the document/text/section; produce a self-contained Q&A

IMPORTANT - DO NOT generate questions about:

- Page numbers, indexes, or table of contents
- Who is mentioned/listed in the document
- Where information appears in the document structure
- Document metadata or navigation elements

Format your response as:

Question: [Your question here]

Answer: [Your answer here]

Example:

Question: What is the budget allocation mentioned for defense procurement?

Answer: \$2.4 billion

EXPLAIN.

Based on the provided context, generate a question that asks for an explanation or clarification of a concept, distinction, or technical detail mentioned in the text.

Context:
{context}

Requirements:

- Question should ask for an explanation, clarification, or distinction (prefer 'How...' or 'Why...' unless a clear 'What is the difference...' is present)
- Answer must be 2-3 sentences:
 - Sentence 1: direct explanation (what/how/why) using terms from the Context.
 - Sentence 2-3: tie the explanation to one or two specific supporting clauses (quote 2-8 words) from the Context.
- Ground the explanation in specific source sentences; avoid generic or speculative claims
- Use only information present in the Context; do not add background not present
- Where helpful, use short phrases from the source for key terms
- Do not refer to the document/text/section; produce a self-contained Q&A

Format your response as:

Question: [Your question here]

Answer: [Your answer here]

Example:

Question: What is the difference between operational capability and strategic capacity?

Answer: Operational capability refers to the immediate ability to conduct specific military operations, while strategic capacity represents the long-term potential to sustain and expand operations over time.

SUMMARIZE.

Based on the provided context, generate a question that asks for a summary or overview of the key points, and provide a concise summary as the answer.

Context:
{context}

Requirements:

- Question should ask for a summary, overview, or key points
- Answer should be a comprehensive but concise summary (2-4 sentences)
- Include the most important information from the context
- Use specific names, figures, and terms where they appear (avoid vague phrasing)
- Structure the summary logically and clearly
 - Include at least one concrete figure or named program if present
 - Cover what is being done (action), why (rationale), and scale (numbers/timeframes) when present
 - Avoid speculation or recommendations; report only what is explicitly stated
 - Use only information present in the Context; do not add external knowledge
 - Do not refer to the document/text/section; produce a self-contained Q&A

Format your response as:

Question: [Your question here]

Answer: [Your answer here]

Example:

Question: What are the key points of the cybersecurity strategy outlined in this section?

Answer: The cybersecurity strategy focuses on three main areas: strengthening defensive capabilities through advanced threat detection systems, building workforce expertise through specialized training programs, and enhancing international cooperation through information sharing agreements. The strategy emphasizes a proactive approach to cyber threats with increased investment in both technology and human resources.

PROVIDE.

Based on the provided context, generate a question that asks for specific quantitative data (budgets, numbers, percentages, dates) and provide the precise numerical answer.

Context:
{context}

Requirements:

- Question should ask for specific quantitative information
- Answer should copy the exact numeric phrase from the context (including units)
- If the number appears within a sentence, quote the key clause containing it
- Include units where appropriate (dollars, percentages, dates, quantities)
- The answer must be factually accurate and directly from the text
- NO HEDGING: Do not use words like "likely", "may", "could", "approximately", "suggests". Provide definitive figures only as stated in the context.
- NO INFERENCE: Do not infer or estimate numbers; only copy what is explicitly present.
- If a range is shown and the question asks for total/planned/overall, copy the range exactly (e.g., "\$5.9bn – \$7.9bn"). If the question asks for 'approved' vs 'unapproved', pick the specific column value.
- If no numeric value exists for your question, select a different quantitative question that can be answered exactly from the Context.
- Use only information present in the Context; do not add external knowledge
- Do not refer to the document/text/section; produce a self-contained Q&A

IMPORTANT - DO NOT generate questions about:

- Page numbers or document navigation
- Index entries or table of contents data
- Which page or section contains information
- Document structure or metadata

Format your response as:

Question: [Your question here]

Answer: [Your answer here]

Example:

Question: What is the total budget allocation for procurement mentioned in the report?

Answer: \$1.2 billion

GENERATE.

Based on the provided context, generate a question that asks for a structured list (capabilities, priorities, requirements, etc.) and provide the answer as a numbered list.

Context:
{context}

Requirements:

- Question should ask for a list of items (capabilities, priorities, gaps, requirements, etc.)
- Answer should be a numbered list with 3-6 items
- Each item must be grounded in a single supporting sentence from the Context
- Include a short quoted anchor (2-6 words) from that sentence for each item
- Do not merge information from multiple sentences into one item
- Use phrases directly from the source sentence (verbatim or near-verbatim)
- Do not introduce items that are not explicitly stated in the Context
- Items should be substantial noun phrases (avoid weak forms like 'there is/are')
- Use only information present in the Context; do not add external knowledge
- Do not refer to the document/text/section; produce a self-contained Q&A

Format your response as:

Question: [Your question here]

Answer: [Your numbered list here]

Example:

Question: What are the key capabilities mentioned for the new defense system?

Answer: 1. Advanced threat detection using AI-powered surveillance systems

2. Rapid response coordination through integrated command networks

3. Multi-domain operational flexibility across land, sea, and air environments

4. Enhanced cybersecurity protection for critical infrastructure