

URGen: Uncertainty-guided Retrieval-augmented Generation for Guideline-Grounded Medical Question Answering

Mengieong Hoi^{1,2} Chen Zhao^{3,4} Chaoyu Lei^{3,4} Huifang Zhou^{3,4} Zhedong Zheng^{1,2*}

¹ Faculty of Information Science and Computing, University of Macau, Macau, China

² Institute of Artificial Intelligence and Brain Sciences, University of Macau, Macau, China

³ State Key Laboratory of Eye Health, Department of Ophthalmology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

⁴ Hainan International Medical Center, Shanghai Jiao Tong University School of Medicine, Hainan, China

{yc67382,zhedongzheng}@um.edu.mo; {Zhaoc1390shsmu,lcy315,fangzzfang}@sjtu.edu.cn

Abstract

Large language models (LLMs) can hallucinate in specialized medical QA and remain poorly calibrated. We propose Uncertainty-guided Retrieval-augmented Generation (URGen), a training-free framework for disease-scoped, guideline-grounded multiple-choice QA. URGen ranks guideline keypoints under multiple temperatures, uses ranking entropy to select retrieval depth, aggregates multi-temperature answers, and fuses normalized context and answer entropy. On TED with Llama-3.1-8B, five-seed evaluation gives $80.40 \pm 1.29\%$ accuracy, $1.98 \pm 0.00\%$ overconfident errors, and $6.80 \pm 1.13\%$ expected calibration error. Paired bootstrap tests show lower overconfident error than every fixed depth ($p \leq 0.0016$), while accuracy remains comparable to Fixed-1 ($p = 0.0630$). A broader descriptive comparison covers two thyroid-related ophthalmology benchmarks and three backbones. The evidence therefore supports URGen primarily as a reliability-oriented method for controlled guideline galleries, rather than a general web-scale retriever.

1 Introduction

Trustworthy medical question answering (QA) can reduce misdiagnosis, promote consistent interpretation of clinical guidelines, and broaden access to specialist knowledge (Jin et al., 2021; Pal et al., 2022; Wang et al., 2025; Rezaei et al., 2025). However, medical questions often depend on long-tail, domain-specific facts that lie outside the pretraining distribution of general-purpose LLMs (Vilares and Gómez-Rodríguez, 2019; Xu et al., 2025). As a result, even strong models usually produce fluent but incorrect answers with high confidence (Li et al., 2023; Simhi et al., 2025), which is unacceptable in safety-critical settings.

*Corresponding author.

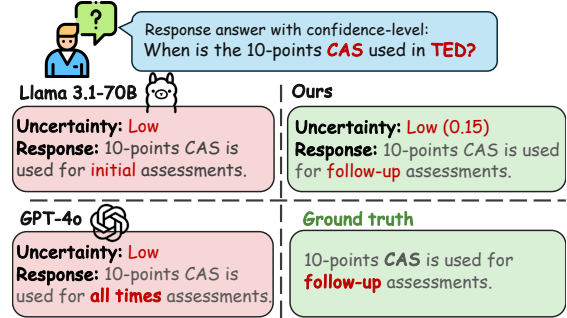


Figure 1: *left*: A motivational example compares outputs from two large language models (LLMs) for the medical question “When is the 10-point CAS used in TED?” Both Llama 3.1-70B (Llama Team, 2024) and GPT-4o-2024-1120 (OpenAI, 2023) confidently return incorrect answers, *i.e.*, *initial* and *all times*, with low uncertainty, *i.e.*, high confidence. This figure illustrates the failure mode rather than comparing the backbones evaluated in Section 4. *right*: Conversely, our method provides the correct answer alongside a quantified uncertainty score. The ground truth answer is *follow-up* assessments.

In Figure 1, we illustrate the failure mode on an ophthalmology guideline question, where both Llama-3.1-70B (Llama Team, 2024) and GPT-4o-2024-1120 (OpenAI, 2023) return hallucinated answers without meaningful semantic-level uncertainty signals (despite token-based or $p(\text{True})$ -style confidence estimates (Kadavath et al., 2022; Tian et al., 2023)). Therefore, reliable medical QA requires (i) grounding answers in authoritative external evidence and (ii) calibrated uncertainty estimates that reflect ambiguity, missing evidence, or conflicts (Guo et al., 2017; Kuleshov et al., 2018; Huang et al., 2024). This improves transparency and overall user-friendliness, fostering greater trust and willingness to use the AI among patients and clinicians.

Existing approaches, *e.g.*, retrieval-augmented generation (RAG), partially address grounding by retrieving top- k passages before generation (Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020; Jiang et al., 2023b; Chen et al., 2024; Hwang

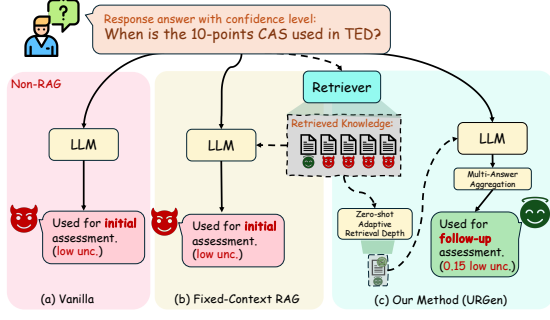


Figure 2: Comparison of inference pipelines for guideline-grounded medical question answering. (a) **Vanilla (Non-RAG)**: The query is directly input into the large language model (LLM), producing a single answer without external evidence or a context-uncertainty signal. (b) **Fixed-Context RAG**: A retriever provides a fixed set of passages that are concatenated with the query before LLM inference. Although retrieval adds context, irrelevant or noisy passages usually persist, and uncertainty remains uncalibrated. (c) **Our Method (URGen)**: Retrieved passages are first filtered by *Zero-shot Adaptive Retrieval Depth* (ZARD) to retain only relevant evidence. The LLM then generates responses under multiple temperatures, and uncertainty is estimated through entropy-based fusion of these outputs. This adaptive retrieval and calibrated aggregation jointly target evidence selection and reliability estimation.

et al., 2025; Choi et al., 2025) (see Figure 2). Yet fixed-context RAG is brittle in specialized domains: large k can introduce irrelevant or even abnormal passages that dilute key evidence, while small k can omit critical guidance (Figure 2) (Izacard and Grave, 2021; Shi et al., 2024). Meanwhile, common uncertainty estimation methods, *e.g.*, predictive entropy, dropout, and temperature scaling, mainly operate at the token level and frequently miss semantic errors or unsafe advice (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016; Guo et al., 2017; Zhang et al., 2024a; Kruse et al., 2025; Shelmanov et al., 2025; Feng et al., 2025).

We propose **Uncertainty-guided Retrieval-augmented Generation (URGen)**, a training-free framework in which context uncertainty connects evidence selection and reliability estimation. URGen (1) performs multi-temperature knowledge retrieval, (2) applies *Zero-shot Adaptive Retrieval Depth* (ZARD) to adjust context size from the entropy of aggregated relevance scores, (3) generates multiple candidate answers under diverse temperatures and aggregates their option labels by majority vote, and (4) fuses context and answer uncertainty. Thus, uncertainty estimation and adaptive retrieval form one pipeline: the ranking distribu-

tion determines the evidence budget and also contributes to the reported uncertainty. Our contributions are:

- We propose **URGen**, a training-free, uncertainty-guided retrieval-augmented generation framework that incorporates entropy-based adaptive retrieval depth (ZARD) and fused uncertainty calibration, tailored for guideline-grounded medical multiple-choice QA.
- On the five-seed TED evaluation, URGen significantly reduces overconfident errors against all fixed depths while maintaining accuracy comparable to Fixed-1; a broader comparison covers three backbones on TED and GD.

2 Related Work

Retrieval-augmented generation and adaptive retrieval. Retrieval-augmented QA commonly combines a retriever with a generator or reader, where dense retrievers improve evidence recall compared to sparse baselines (Karpukhin et al., 2020). Generative readers such as Fusion-in-Decoder (FiD) condition on multiple retrieved passages and fuse evidence in the decoder, enabling strong performance on knowledge-intensive QA (Izacard and Grave, 2021). Subsequent work improves robustness by inserting reranking between retrieval and generation (Mao et al., 2021; Glass et al., 2022) and by supporting black-box or medical RAG pipelines via plug-and-play retrieval interfaces and evidence-aware retrieval (Shi et al., 2024; Liang et al., 2025; Sohn et al., 2025; Chen et al., 2024; Hwang et al., 2025; Choi et al., 2025; Xu et al., 2025). A key practical issue is that fixed retrieval depth can be brittle: small depth misses necessary evidence, while large depth introduces distractors. Prior work addresses this via learned or iterative information-seeking policies that dynamically choose retrieval actions or stop early (Zhu et al., 2021; Jiang et al., 2023b). Our work follows this line but targets a training-free, uncertainty-guided depth adaptation for guideline-grounded medical multiple-choice QA.

Uncertainty estimation and calibration. Reliable uncertainty is central for safe deployment, yet modern LMs are often miscalibrated (Desai and Durrett, 2020; Tian et al., 2023; Huang et al., 2024; Zhang et al., 2024a; Kruse et al., 2025; Shelmanov

et al., 2025; Feng et al., 2025). For QA, calibration has been studied for language models and prompting pipelines, including how confidence behaves under in-context learning (Jiang et al., 2021; Zhang et al., 2024b; Tian et al., 2023). Semantic entropy clusters sampled generations by meaning before measuring their dispersion (Kuhn et al., 2023); retrieval-aware approaches further examine whether uncertainty changes with the available evidence (Perez-Beltrachini and Lapata, 2025). In our constrained multiple-choice setting, option labels provide the answer categories, so we use Shannon entropy over the aggregated option distribution rather than proposing a new semantic-clustering estimator. Softmax normalization, entropy, multi-temperature sampling, and majority voting are established components. Our contribution is their coupling: ranking entropy controls retrieval depth, and the resulting context uncertainty is fused with answer uncertainty to quantify reliability.

Medical QA benchmarks. Medical QA spans clinical notes, biomedical literature, and exam-style multiple-choice questions. Representative resources include emrQA for EMR question answering (Pampari et al., 2018), PubMedQA for research questions (Jin et al., 2019), MedNLI for clinical inference (Romanov and Shivade, 2018), and MEDIQA shared tasks for medical QA and entailment (Ben Abacha et al., 2019). Exam-oriented multiple-choice benchmarks such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and HEAD-QA (Vilares and Gómez-Rodríguez, 2019; Wang et al., 2025; Rezaei et al., 2025) are increasingly used to evaluate medical reasoning under constrained answer options, where calibration and retrieval quality both directly affect safety.

3 Methodology

3.1 Framework Overview

As shown in Figure 3, we introduce Uncertainty-guided Retrieval-augmented Generation (URGen), a training-free framework that harnesses an off-the-shelf frozen large language model (LLM) for guideline-grounded medical multiple-choice QA. We do not pursue the best possible backbone, but instead focus on mitigating hallucinations and improving reliability in specialized domains. URGen consists of four stages: multi-temperature knowledge retrieval, zero-shot adaptive retrieval

depth (ZARD), multi-answer aggregation, and uncertainty estimation.

We denote q as a multiple-choice medical question with options $\mathcal{A}$ and ground-truth choice $\tilde{a} \in \mathcal{A}$. We assume access to a domain knowledge corpus $\mathcal{C} = \{(k_i; t_i)\}_{i=1}^N$ constructed from expert resources and clinical consensus documents, where each k_i is a summarized keypoint and t_i is the corresponding passage. Our goal is to learn a mapping function:

$$f : (q, \mathcal{C}) \longrightarrow (\hat{a}, \hat{u}), \quad (1)$$

where $\hat{a}$ denotes the predicted option and $\hat{u} \in [0, 1]$ is a calibrated uncertainty score for the prediction.

3.2 Multi-Temperature Knowledge Retrieval

First, we parse clinical guidelines and consensus documents into a structured collection of keypoint–passage pairs $\mathcal{C} = \{(k_i; t_i)\}$. As detailed in Appendix A (Table S2), each pair comprises a guideline statement and its supporting section, extracted deterministically from the public XML structure. Ophthalmology experts authored and verified the QA items and labels. We do not evaluate learned or noisy automatic extraction in this study. Here, k_i is a concise statement capturing the main claim or finding of a passage, and t_i is the full text content of that passage. Due to the lengthy full content, we perform retrieval using the question q and keypoint collection $\mathcal{K} = \{k_i\}$ instead of the full content $\mathcal{T} = \{t_i\}$ as query. Given a question q , the keypoint collection $\mathcal{K}$ and a prompt tailored for ranking keypoints, we query a frozen LLM M_1 times with different temperatures $\{\tau_1, \dots, \tau_{M_1}\}$ to generate M_1 ranking lists $\{l_1, \dots, l_{M_1}\}$ of keypoints in descending order of relevance. This multi-temperature retrieval provides multiple “views” of relevance that later support adaptive depth selection. The detailed ranking prompt is provided in Appendix G.

3.3 Zero-Shot Adaptive Retrieval Depth

We assign a positive relevance score to each of the top- K ($K \leq N$) keypoints in the ranking list based on their ranking position. For each unique keypoint k_i , we aggregate the relevance score s_i by summing its raw scores across all M_1 ranking lists l and applying a softmax normalization as:

$$s_i = \frac{\exp(\sum_{j=1}^{M_1} \text{score}_i^{(l_j)})}{\sum_{i=1}^N \exp(\sum_{j=1}^{M_1} \text{score}_i^{(l_j)})}. \quad (2)$$

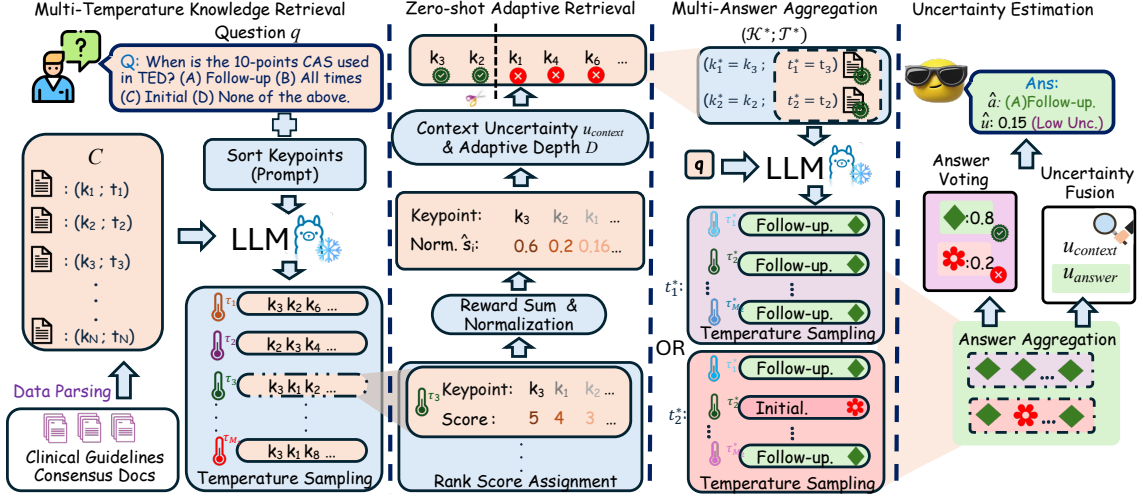


Figure 3: **Overview of the URGen pipeline for guideline-grounded medical question answering.** (1) We parse clinical guidelines and consensus documents into structured keypoint–passage pairs $(k; t)$ for retrieval. Given a query, a *frozen* LLM ranks all keypoints under multiple sampling temperatures, producing diverse orderings. (2) Each ordering is converted into softmax-normalized relevance scores, which are aggregated into a unified relevance distribution. We compute context uncertainty from this distribution and define its effective support. A broader distribution retrieves more passages, whereas a sharply peaked distribution retrieves fewer. (3) The selected passages are concatenated with the query and passed to the *frozen* LLM under multiple answer-generation temperatures, producing candidate answers. (4) We aggregate the predicted options and calculate answer uncertainty from their vote distribution. The final uncertainty fuses normalized context and answer uncertainty. Finally, the pipeline outputs a final prediction $\hat{a}$ accompanied by a calibrated uncertainty estimate $\hat{u}$.

Now, given multiple ranking lists, we have an aggregated relevance score s_i for every keypoint. The softmax does not change the ordering induced by the summed rank scores; it places them on a common probability scale so that aggregation and entropy are well defined across questions. After retaining the top K , power normalization sharpens and renormalizes this restricted distribution. We only take the top K keypoints with large s_i into the next-step consideration. To sharpen the score distribution of these top K keypoints, a power normalization with hyper-parameter α is further applied:

$$\hat{s}_i = \frac{s_i^\alpha}{\sum_{i=1}^K s_i^\alpha}. \quad (3)$$

We quantify context uncertainty by the entropy of the sharpened relevance distribution. Intuitively, when relevance is concentrated on a few keypoints, the entropy is low and we select a compact context; when relevance is diffuse, entropy is higher and we retrieve a broader set to reduce the risk of missing key evidence. The raw context uncertainty u_{context} is the information entropy of the sharpened distribution $\{\hat{s}_i\}$:

$$u_{\text{context}} = - \sum_{i=1}^K \hat{s}_i \log(\hat{s}_i). \quad (4)$$

This measure reflects the concentration of the ranking distribution: low entropy indicates concentrated relevance, whereas high entropy indicates that several keypoints receive similar support. Inspired by the uncertainty-to-scale parameterization in Bayesian uncertainty modeling (Kendall and Gal, 2017), we define $\exp(u_{\text{context}})$ as the effective support and determine the retrieval depth as:

$$D = \text{clip}_{[1,5]}(\lfloor \exp(u_{\text{context}}) \rfloor), \quad (5)$$

and select the top D keypoints according to the descending relevance score s_i as the crucial keypoints $\mathcal{K}^* = \{k_1^*, \dots, k_D^*\}$. The corresponding full contexts $\mathcal{T}^* = \{t_1^*, \dots, t_D^*\}$ form the context for question answering. The exponential and linear mappings are compared in Appendix C. Note that the subscript in k_i^* indicates the relevance rank, with $i = 1$ denoting the highest-ranked keypoint, rather than its corpus index in k_i .

3.4 Multi-Answer Aggregation

We apply zero-shot adaptive retrieval depth (ZARD) to select a limited set of passages $\mathcal{T}^*$, producing a compact input under the evaluated depth bound. In practice, we concatenate the selected passages in $\mathcal{T}^*$ with the query q as augmented context. We then invoke the LLM under M_2 temper-

atures $\{\tau_1^*, \dots, \tau_{M_2}^*\}$, obtaining one predicted option label $\hat{a}^{(j)}$ from each generation. Since the task is multiple-choice, we form the option-vote distribution

$$p(a) = \frac{1}{M_2} \left| \left\{ j : \hat{a}^{(j)} = a \right\} \right|, \quad (6)$$

where $a \in \mathcal{A}$ and $\mathcal{A}$ is the multiple-choice answer set. The final answer is selected by majority vote:

$$\hat{a} = \arg \max_{a \in \mathcal{A}} p(a). \quad (7)$$

3.5 Uncertainty Estimation

For the final uncertainty estimation, we consider two complementary sources: context uncertainty and answer uncertainty. Context uncertainty u_{context} reflects the concentration of the retrieved-keypoint distribution (Eq. 4); it is not a direct measure of clinical or disease uncertainty. In contrast, answer uncertainty u_{answer} captures dispersion in the aggregated option distribution. We define the raw answer entropy as:

$$u_{\text{answer}} = - \sum_{a \in \mathcal{A}} p(a) \log p(a). \quad (8)$$

For fusion only, both entropies are normalized to $[0, 1]$:

$$\tilde{u}_{\text{context}} = \frac{u_{\text{context}}}{\log K}, \quad \tilde{u}_{\text{answer}} = \frac{u_{\text{answer}}}{\log |\mathcal{A}_q|}, \quad (9)$$

where $K = 5$ and $|\mathcal{A}_q|$ is the number of choices for question q . To obtain the overall uncertainty score, we fuse the normalized values as follows:

$$\hat{u} = \frac{\tilde{u}_{\text{answer}} \tilde{u}_{\text{context}}}{\tilde{u}_{\text{answer}} \tilde{u}_{\text{context}} + (1 - \tilde{u}_{\text{answer}})(1 - \tilde{u}_{\text{context}})}. \quad (10)$$

The observed normalized ranges on both datasets are reported in Appendix C. The fused score is sensitive to both retrieval ambiguity and answer-generation dispersion without treating ranking entropy as proof of missing evidence. Ultimately, our method outputs the predicted answer $\hat{a}$ along with the calibrated uncertainty $\hat{u}$.

4 Experiment

4.1 Datasets

We evaluate URGen on two expert-curated ophthalmology multiple-choice datasets, *i.e.*, Thyroid Eye Disease (TED) and Graves’ Disease (GD). Unlike broad medical QA benchmarks such as MedQA and MedMCQA, which mainly assess parametric medical knowledge across general

medical topics, our datasets pair expert-authored questions and labels with authoritative guideline passages. We focus on thyroid-related ophthalmology because its guidelines contain clinically adjacent recommendations whose distinctions depend on disease activity, severity, treatment stage, and imaging or surgical context. This provides a controlled test of evidence selection, but it does not establish cross-specialty generality. The keypoint–passage pairs are extracted deterministically from the public guideline XML structure. Ophthalmology experts participated in question design and answer verification. We will release the corpus to support reproducibility and future research on guideline-grounded medical RAG. Specifically, TED contains 101 questions and a complete disease-scoped gallery of 35 keypoint–passage pairs from *Management of Thyroid Eye Disease by the American and European Thyroid Associations* (Burch et al., 2022); GD contains 97 questions and 22 pairs from the *2018 European Thyroid Association Guideline for the Management of Graves’ Hyperthyroidism* (Kahaly et al., 2018). These are controlled guideline galleries, not web-scale retrieval corpora.

4.2 Compared Methods

We compare our method with three widely-used strategies, *i.e.*, Vanilla, *Fixed-1* RAG, and *Fixed-5* RAG based on prevailing LLMs:

Vanilla. The model receives only the question q with the instruction “*Select the best answer for the question.*” No external context or uncertainty estimation is applied. This configuration measures the LLM’s zero-shot recall of medical knowledge.

Fixed- k RAG. First, we retrieve the top- k keypoint–passage pairs from $\mathcal{C}$ using multi-temperature knowledge retrieval and concatenate them with q in descending relevance order. Second, we sample the LLM under multiple answer-generation temperatures. Context size is fixed at k , and no context-uncertainty calibration is applied.

4.3 Implementation Details

All experiments adopt the same retrieval and inference infrastructure, as detailed below: **Large Language Models (LLMs).** We evaluate three prevailing pretrained backbones: Llama-3.1-8B with BF16 (Llama Team, 2024) and Mistral-7B v0.3 with BF16 (Jiang et al., 2023a) (general-domain); Meditron-7B with BF16 (Chen et al., 2023) (medical fine-tuned). Open-source models

Table 1: Cross-backbone comparison on TED and GD. We report accuracy, OC@0.8, and ECE for three LLM backbones and four inference methods.

Model	Method	Accuracy (%) $\uparrow$		OC@0.8 (%) $\downarrow$		ECE (%) $\downarrow$	
		TED	GD	TED	GD	TED	GD
Llama-3.1-8B	Vanilla LLM	70.30%	46.39%	9.90%	22.68%	10.89%	21.09%
	Fixed-1 RAG	74.65%	42.27%	11.88%	21.65%	14.95%	20.33%
	Fixed-5 RAG	69.70%	50.52%	13.66%	5.15%	19.06%	13.59%
	URGen (Ours)	80.40%	51.55%	1.98%	4.12%	6.80%	9.96%
Mistral-7B	Vanilla LLM	68.32%	44.30%	29.70%	55.67%	24.40%	55.30%
	Fixed-1 RAG	73.26%	48.45%	25.74%	36.08%	6.50%	10.97%
	Fixed-5 RAG	72.28%	47.42%	5.94%	30.93%	10.61%	12.63%
	URGen (Ours)	79.21%	50.52%	3.96%	14.43%	5.31%	8.23%
Meditron-7B	Vanilla LLM	78.22%	63.92%	14.85%	23.71%	12.06%	23.15%
	Fixed-1 RAG	80.20%	61.86%	6.93%	11.34%	3.01%	8.82%
	Fixed-5 RAG	75.25%	59.79%	2.97%	4.12%	11.32%	7.96%
	URGen (Ours)	84.54%	65.98%	1.98%	3.09%	3.01%	4.57%

are loaded via the Hugging Face Transformers library on an NVIDIA 4090 GPU (24 GB memory). **Temperature Settings.** *Retrieval:* $M_1 = 5$ temperatures $\tau_j \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. *Answer Generation:* $M_2 = 5$ temperatures $\tau_j^* \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. Joint sensitivity to three, five, and seven temperatures is reported in Appendix C. **Context Size Bounds.** We retain the top $K = 5$ candidates when forming the ranking distribution and set $(D_{\min}, D_{\max}) = (1, 5)$ for the ZARD retrieval depth.

4.4 Evaluation Metrics

We evaluate (i) correctness and (ii) reliability of the reported uncertainty. Let $|Q|$ be the number of questions, $\hat{a}_i$ the predicted option for q_i , and $y_i \in \{0, 1\}$ the correctness indicator, where $y_i = 1$ if $\hat{a}_i = a_i$ and 0 otherwise. Let $\hat{u}_i \in [0, 1]$ denote the predicted uncertainty, and define confidence as $\hat{c}_i = 1 - \hat{u}_i$. **Accuracy.** $\text{Acc} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} y_i$. **Overconfident error rate (OC@ ζ).** Following overconfidence-focused reliability evaluation (Groot and Valdenegro-Toro, 2024), we report OC@0.8 (*i.e.*, $\zeta = 0.8$) as the fraction of all questions that are answered incorrectly with confidence at least 0.8:

$$\text{OC}(\zeta) = \frac{1}{|Q|} |\{i : y_i = 0, \hat{c}_i \geq \zeta\}|. \quad (11)$$

Expected calibration error (ECE). Following Guo et al. (Guo et al., 2017), we bin predictions by confidence into B equal-width bins. Define the index sets: $\mathcal{I}_b = \{i : \hat{c}_i \in ((b-1)/B, b/B]\}$, $b = 1, \dots, B$. For each non-

empty bin,

$$\text{acc}_b = \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} y_i, \quad \text{conf}_b = \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} \hat{c}_i, \quad (12)$$

and we set $B = 5$:

$$\text{ECE} = \sum_{b: |\mathcal{I}_b| > 0} \frac{|\mathcal{I}_b|}{|Q|} |\text{acc}_b - \text{conf}_b|. \quad (13)$$

OC@0.8 and ECE describe different reliability failures. OC@0.8 isolates high-confidence mistakes and uses all questions as its denominator; ECE summarizes the confidence–accuracy gap across five bins. Because ECE depends on binning and the benchmarks contain about 100 questions each, we interpret it together with accuracy and OC@0.8 rather than as a standalone reliability claim.

Statistical protocol. For the TED/Llama-3.1-8B fixed-depth comparison in Table 3, we use seeds $\{0, 1, 2, 3, 4\}$ and report mean $\pm$ standard deviation. We obtain 95% confidence intervals and two-sided p -values from 10,000 paired question-level bootstrap samples, resampling the 101 questions while retaining all five outcomes per question. Inferential claims refer to this analysis; Table 1 provides the broader cross-backbone comparison.

4.5 Comparison with Competitive Methods

Table 1 compares three backbones on TED and GD. URGen has the highest accuracy and lowest OC@0.8 in all six model–dataset blocks; its ECE is lowest or tied for lowest in all six. On TED with Llama-3.1, URGen obtains 80.40% accuracy, 1.98% OC@0.8, and 6.80% ECE, compared with 74.65%/11.88%/14.95% for Fixed-1

Table 2: Component-wise ablation on TED using Llama-3.1. *: Fixed-1 RAG without multi-temperature sampling in retrieval or answer generation.

Method	Acc. (%) $\uparrow$	OC@0.8 (%) $\downarrow$	ECE (%) $\downarrow$
URGen (full)	80.40%	1.98%	6.80%
w/o uncertainty calibration	80.40%	6.93%	12.31%
w/o adaptive retrieval (<i>Fixed-k</i> = 5)	69.70%	13.66%	19.06%
w/o retrieval temperature diversity ($M_1 = 1$)	52.48%	23.76%	25.37%
w/o generation temperature diversity ($M_2 = 1$)	55.45%	17.82%	19.51%
w/o multi retrieval (<i>Fixed-k</i> = 1)*	71.29%	12.87%	15.36%
Vanilla LLM	70.30%	9.90%	10.89%

Table 3: Five-seed comparison of fixed retrieval depth k versus URGen on TED with Llama-3.1-8B. Values are mean $\pm$ standard deviation.

Method	Acc.(%) $\uparrow$	OC@0.8(%) $\downarrow$	ECE(%) $\downarrow$
<i>Fixed-1</i> RAG	74.65 $\pm$ 1.13	11.88 $\pm$ 1.40	14.95 $\pm$ 2.29
<i>Fixed-2</i> RAG	73.07 $\pm$ 2.47	9.31 $\pm$ 3.02	14.07 $\pm$ 2.44
<i>Fixed-5</i> RAG	69.70 $\pm$ 1.50	13.66 $\pm$ 2.66	19.06 $\pm$ 2.36
URGen	80.40 $\pm$ 1.29	1.98 $\pm$ 0.00	6.80 $\pm$ 1.13

and 69.70%/13.66%/19.06% for Fixed-5. The paired analysis for this setting appears in Table 3 and Appendix C.

5 Ablation Studies

Component removal. Table 2 isolates the main components of URGen. Removing fused calibration leaves accuracy unchanged at 80.40%, while OC@0.8 rises from 1.98% to 6.93% and ECE from 6.80% to 12.31%. Accuracy is unchanged because this ablation preserves retrieval and answer voting and changes only the reported uncertainty. Replacing adaptive retrieval with Fixed-5 reduces accuracy to 69.70% and increases OC@0.8 to 13.66%. Removing either retrieval- or generation-temperature diversity produces the largest accuracy losses. Thus, adaptive depth and temperature diversity primarily support correctness, whereas fusing context and answer uncertainty improves reliability. The fixed-depth comparison is reported in Table 3, and temperature-count sensitivity appears in Appendix C.

Retrieval depth. Table 3 reports the five-seed fixed-depth comparison. URGen has the highest mean accuracy and lowest OC@0.8 among the evaluated depths. In 10,000 paired question-level bootstrap samples, its accuracy difference is significant against Fixed-2 ($p = 0.0318$) and Fixed-5 ($p = 0.0052$), but not Fixed-1 ($p = 0.0630$). OC@0.8 is significantly lower in all three comparisons ($p \leq 0.0016$), whereas the ECE difference is significant only against Fixed-5 ($p = 0.0110$). Appendix Table S5 reports every effect size, confidence interval, and p -value.

Effect of sharpen parameter α . We investigate

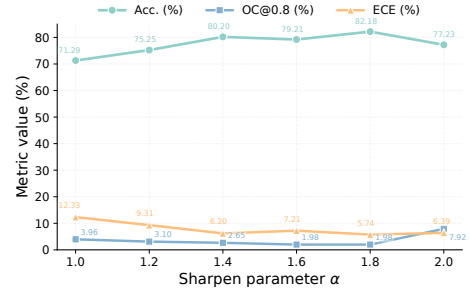


Figure 4: Sensitivity study of the sharpen parameter α using Llama-3.1 on TED. Performance remains stable for $\alpha \in [1.4, 1.8]$, indicating that the observed result is not tied to a single sharpening value.

the sensitivity of retrieval sharpening by varying α from 1.0 to 2.0. As shown in Fig. 4, performance is stable across a broad moderate range. For $\alpha \in [1.4, 1.8]$, accuracy ranges from 79.21% to 82.18%, OC@0.8 from 1.98% to 2.65%, and ECE from 5.74% to 7.21%. The result is therefore not tied to one narrowly tuned α . Lower values provide insufficient discrimination among keypoints, while overly large sharpening may make evidence selection too narrow.

Case Study. We present one TED QA case in Table 4 to illustrate adaptive evidence selection and the potential effect of redundant context. For a question about the most comprehensive TED assessment system, vanilla prompting and Fixed-5 RAG produce incorrect answers. URGen retains four passages centered on activity, severity, and assessment, and correctly identifies VISA. This example illustrates how a compact evidence set can preserve the decisive guideline information, but it is qualitative and does not by itself establish a causal effect of retrieval depth.

Calibration Visualization. As shown in Figure 5, we visualize the distribution of predicted uncertainty for correct versus incorrect answers on TED. Correct answers are concentrated in low-uncertainty bins, whereas errors occur more frequently at higher uncertainty scores.

Sensitivity, distractors, and cost. The additional analyses in Appendix C test the main design

Table 4: We show a typical case comparison on TED among vanilla prompting, fixed-5 RAG, and URGen. We list the question and choices, ground-truth answer, and each method’s prediction with its uncertainty (green for correct, red for incorrect). URGen’s selected keypoints and their ranking scores illustrate its adaptive retrieval mechanism.

Case Study 1	
Model / Corpus	Llama-3.1 8B / TED
Question & Choices	The comprehensive assessment system for evaluating both the activity and severity of thyroid eye disease (TED) is: A. CAS B. NOSPECS C. GO-QOL D. VISA
GT	D. VISA
Vanilla	Answer: B. NOSPECS (Unc.: 7.31%)
<i>fixed-5 RAG</i>	URGen
Answer: A. CAS (Unc.: 35.44%)	Answer: D. VISA (Unc.: 48.94%)
Selected 5 Keypoints	Selected 4 Keypoints
2.1 Activity and severity definitions for patients with TED, CAS score definition. 4.1.1 Endocrinologists should be familiar with basic elements of a TED examination enabling assessment of both activity and severity. 7.1.1.1 Intravenous glucocorticoid (IVGC) therapy is preferred for active moderate-to-severe TED when activity predominates. 4.1.2 Assessment of TED should include activity, severity, trend over time, and impact on daily living. 8.3.1 In compressive DON, orbital decompression of the deep medial wall and floor should be considered.	2.1 Activity and severity definitions for patients with TED, CAS score definition. (Ranking Score: 0.31) 4.1.1 Endocrinologists should be familiar with basic elements of a TED examination enabling assessment of both activity and severity. (Ranking Score: 0.18) 7.1.1.1 Intravenous glucocorticoid (IVGC) therapy is preferred for active moderate-to-severe TED when activity predominates. (Ranking Score: 0.19) 4.1.2 Assessment of TED should include activity, severity, trend over time, and impact on daily living. (Ranking Score: 0.17)

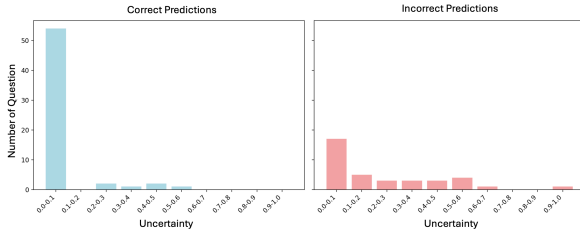


Figure 5: Uncertainty distribution on TED. We plot the number of questions across uncertainty bins (0–1, bin width 0.1) for correct and incorrect predictions. Correct answers (*left*) cluster at low uncertainty, while incorrect answers (*right*) occur more frequently at higher uncertainty.

choices. Jointly varying the retrieval and answer temperature counts shows that $(M_1, M_2) = (5, 5)$ gives the highest mean accuracy (80.40%) and lowest OC@0.8 (1.98%), whereas seven temperatures give lower ECE (4.00%) at greater sampling cost. The five-temperature setting therefore balances correctness and overconfident errors; it is not uniformly best on every metric. Replacing the effective-support mapping in Eq. 5 with a normalized linear mapping yields 78.81% accuracy and 7.39% OC@0.8, compared with 80.40% and 1.98% for the exponential mapping, while ECE remains similar (6.60% versus 6.80%).

Adding all 22 GD keypoints to the 35-keypoint TED gallery lowers every method, but URGen remains strongest at 77.03%, versus 72.48% for Fixed-1 and 68.91% for Fixed-5 (Appendix Table S8). This is a disease-specific distractor test, not evidence of web-scale scalability. Appendix C also reports the normalized uncertainty ranges and shows that 2 of the 26 questions in the lowest context-uncertainty quartile are answered in-

correctly. The measured runtime is 4.33 GPU-s/question for URGen, between Fixed-1 (3.78) and Fixed-5 (5.48).

Evaluation scope. Table 1 tests consistency across backbones, while Table 3 and the paired bootstrap provide focused reliability evidence. The ablations, sensitivity analyses, distractor test, and runtime study characterize the method’s behavior and cost. Accordingly, Table 1 supports a descriptive consistency claim rather than statistical significance for every model–dataset block. The evidence is limited to curated guideline galleries; automatic and web-scale retrieval remain outside the evaluation.

6 Conclusion

We introduce URGen, a training-free framework that uses context uncertainty to adapt retrieval depth and complement answer uncertainty in guideline-grounded medical multiple-choice QA. On TED with Llama-3.1-8B, URGen significantly reduces overconfident errors against every fixed depth while maintaining accuracy comparable to Fixed-1. Cross-backbone and sensitivity results further support the method within curated disease-scoped galleries. URGen’s uncertainty is a triage signal rather than a correctness guarantee. Future work should evaluate automatically extracted corpora, additional specialties, and open-ended questions.

Limitations

URGen assumes that a flatter keypoint-ranking distribution warrants more context. It may fail when the model is consistently confident about wrong or incomplete keypoints: ZARD can then assign low entropy, shrink the context, and miss necessary evidence. Appendix Table S10 quantifies this failure mode: 2 of 26 errors occur in the lowest uncertainty quartile and 5 of 51 in the lower half. These counts do not directly establish ranking correctness because gold keypoint labels are unavailable. URGen also cannot recover evidence absent from its corpus. Our evaluation uses complete disease-scoped galleries of 35 TED or 22 GD keypoints; the pooled-gallery test in Appendix Table S8 adds disease-specific distractors but does not establish robustness for web-scale retrieval or automatically extracted, noisy corpora.

Computationally, URGen’s multi-temperature ranking and answer aggregation increase total inference work. Appendix Table S3 reports 4.33 GPU-s/question, above Fixed-1 (3.78) but below Fixed-5 (5.48); batching independent branches does not remove this total-compute cost. Finally, the present evidence is limited to multiple-choice ophthalmology QA with authoritative guidelines. Applying URGen to other specialties requires a new verified corpus, and its behavior for open-ended tasks remains untested.

Acknowledgments

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515012281), the Nanjing Municipal Science and Technology Bureau (Grant No. 202401035), the Macao Science and Technology Development Fund (Grant No. 0209/2025/AMJ), the National Natural Science Foundation of China (Grant Nos. 82388101 and 82271122), the “Explorer Program” (First Batch) of the 2026 Basic Research Plan (Grant No. 26TS1401900), the National Key R&D Program of China (Grant Nos. 2024YFB4710200 and 2024YFB4710205), the Shanghai Jiao Tong University 2030 Initiative (Grant No. WH510272301), the Science and Technology Commission of Shanghai Municipality (Grant No. 23JC1402402), the Shanghai Three-Year Plan for the Inheritance and Innovative Development of Traditional Chinese Medicine (Grant No. 2-1-5), the National Administration of Traditional Chinese Medicine

Clinical Collaboration Program for Major and Difficult Diseases (Grant No. ZDYN-2024-A-091), the Hainan Provincial Key Research and Development Projects (Grant No. ZDYF2024LCLH004), and the Shanghai Jiao Tong University School of Medicine Vertical Research Projects (Grant Nos. yskt2-24-0926-2 and yskfl-24-0926-2).

References

- Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. [Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.
- Henry B. Burch, Petros Perros, Tomasz Bednarczuk, David S. Cooper, Peter J. Dolman, Angela M. Leung, Ilse Mommaerts, Mario Salvi, and Marius N. Stan. 2022. [Management of thyroid eye disease: A consensus statement by the american thyroid association and the european thyroid association](#). *European Thyroid Journal*, 11(6).
- Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. [RQ-RAG: Learning to refine queries for retrieval augmented generation](#). In *Proceedings of the First Conference on Language Modeling*.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. [Reading wikipedia to answer open-domain questions](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879. Association for Computational Linguistics.
- Lu Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2024. [Controlling risk of retrieval-augmented generation: A counterfactual prompting framework](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2380–2393, Miami, Florida, USA. Association for Computational Linguistics.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. [MEDITRON-70b: Scaling medical pretraining for large language models](#). *arXiv preprint arXiv:2311.16079*.
- Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. [Conflict-aware soft prompting for retrieval-augmented generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26981–26995,

- Suzhou, China. Association for Computational Linguistics.
- Shrey Desai and Greg Durrett. 2020. [Calibration of pre-trained transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302. Association for Computational Linguistics.
- Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishalay Halder, Yang Li, Yassine Benajiba, and Dan Roth. 2025. [Rethinking llm uncertainty: A multi-agent approach to estimating black-box model uncertainty](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12349–12375, Suzhou, China. Association for Computational Linguistics.
- Yarin Gal and Zoubin Ghahramani. 2016. [Dropout as a bayesian approximation: Representing model uncertainty in deep learning](#). In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR.
- Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. [Re2G: Retrieve, rerank, generate](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.
- Tobias Groot and Matias Valdenegro-Toro. 2024. [Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models](#). In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing*, pages 145–171. Association for Computational Linguistics.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Papat, and Ming-Wei Chang. 2020. [Retrieval augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938. PMLR.
- Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. [Uncertainty in language models: Assessment through rank-calibration](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312, Miami, Florida, USA. Association for Computational Linguistics.
- Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. [Retrieval-augmented generation with estimation of source reliability](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34279–34303, Suzhou, China. Association for Computational Linguistics.
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023a. [Mistral 7b](#). *arXiv preprint arXiv:2310.06825*.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. [Active retrieval augmented generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2567–2577. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislaw Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *arXiv preprint arXiv:2207.05221*.
- George J. Kahaly, Luigi Bartalena, Laszlo Heged  us, Laurence Leenhardt, Kris Poppe, and Simon H. Pearce. 2018. [2018 european thyroid association guideline for the management of graves’ hyperthyroidism](#). *European Thyroid Journal*, 7(4):167–186.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and

- Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.
- Alex Kendall and Yarin Gal. 2017. [What uncertainties do we need in bayesian deep learning for computer vision?](#) In *Advances in Neural Information Processing Systems*, volume 30.
- Maya Kruse, Majid Afshar, Saksham Khatwani, Anoop Mayampurath, Guanhua Chen, and Yanjun Gao. 2025. [Simple yet effective: An information-theoretic approach to multi-llm uncertainty quantification](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30493–30504, Suzhou, China. Association for Computational Linguistics.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *International Conference on Learning Representations*.
- Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. [Accurate uncertainties for deep learning using calibrated regression](#). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2796–2804. PMLR.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. [Simple and scalable predictive uncertainty estimation using deep ensembles](#). In *Advances in Neural Information Processing Systems*, volume 30.
- 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.
- Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. [HaluEval: A large-scale hallucination evaluation benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.
- Sichu Liang, Linhai Zhang, Hongyu Zhu, Wenwen Wang, Yulan He, and Deyu Zhou. 2025. [Rgar: Recurrence generation-augmented retrieval for factual-aware medical question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4006–4033, Suzhou, China. Association for Computational Linguistics.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. [Generating with confidence: Uncertainty quantification for black-box large language models](#). *Transactions on Machine Learning Research*.
- Llama Team. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. [Reader-guided passage reranking for open-domain question answering](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 344–350. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. [MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260. PMLR.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. [emrQA: A large corpus for question answering on electronic medical records](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368. Association for Computational Linguistics.
- Laura Perez-Beltrachini and Mirella Lapata. 2025. [Uncertainty quantification in retrieval augmented question answering](#). *Transactions of Machine Learning Research*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Mohammad R. Rezaei, Reza Saadati Fard, Jayson L. Parker, Rahul G. Krishnan, and Milad Lankarany. 2025. [Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between LLMs and evolving medical knowledge](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12682–12701, Suzhou, China. Association for Computational Linguistics.
- Alexey Romanov and Chaitanya Shivade. 2018. [Lessons from natural language inference in the clinical domain](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596. Association for Computational Linguistics.
- Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. 2025.

- A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in llm outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35712–35731, Suzhou, China. Association for Computational Linguistics.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. **REPLUG: Retrieval-augmented black-box language models**. 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 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.
- Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. **Trust me, I’m wrong: LLMs hallucinate with certainty despite knowing the answer**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14665–14688, Suzhou, China. Association for Computational Linguistics.
- Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. **Rationale-guided retrieval augmented generation for medical question answering**. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 12739–12753, Albuquerque, New Mexico. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. **Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- David Vilares and Carlos Gómez-Rodríguez. 2019. **HEAD-QA: A healthcare dataset for complex reasoning**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966. Association for Computational Linguistics.
- Yinuo Wang, Baiyang Wang, Robert E. Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. 2025. **Trustworthy medical question answering: An evaluation-centric survey**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27489–27502, Suzhou, China. Association for Computational Linguistics.
- Kaishuai Xu, Wenjun Hou, Yi Cheng, and Wenjie Li. 2025. **RAR²: Retrieval-augmented medical reasoning via thought-driven retrieval**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20386–20396, Suzhou, China. Association for Computational Linguistics.
- Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. **LUQ: Long-text uncertainty quantification for LLMs**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.
- Hanlin Zhang, YiFan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Himabindu Lakkaraju, and Sham Kakade. 2024b. **A study on the calibration of in-context learning**. 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 6118–6136, Mexico City, Mexico. Association for Computational Linguistics.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, and 5 others. 2025. **A multimodal biomedical foundation model trained from fifteen million image–text pairs**. *NEJM AI*, 2(1).
- Yunchang Zhu, Liang Pang, Yanyan Lan, Huawei Shen, and Xueqi Cheng. 2021. **Adaptive information seeking for open-domain question answering**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3626. Association for Computational Linguistics.

URGen: Uncertainty-guided Retrieval-augmented Generation for Guideline-Grounded Medical Question Answering

Supplementary Materials

A Dataset and Keypoint Corpus Construction

Dataset overview. We construct two expert-curated ophthalmology multiple-choice QA benchmarks from long-form clinical guidelines and consensus documents: Thyroid Eye Disease (TED) and Graves' Disease (GD). The TED benchmark is based on the *Management of Thyroid Eye Disease by the American and European Thyroid Associations*¹ (Burch et al., 2022), while the GD benchmark is based on the *2018 European Thyroid Association Guideline for the Management of Graves' Hyperthyroidism*² (Kahaly et al., 2018). These documents are substantially longer and more specialized than the passages typically used in general-domain reading comprehension. They contain diagnostic criteria, activity and severity definitions, treatment recommendations, contraindications, referral criteria, and follow-up guidance, all of which are important for clinically grounded medical QA.

The resulting datasets are designed to support several evaluation goals. First, they evaluate whether a model can answer specialized medical questions that require guideline-level knowledge rather than broad medical commonsense. Second, because every question is paired with a curated keypoint corpus, the datasets allow controlled evaluation of retrieval quality, adaptive retrieval depth, and evidence sufficiency. Third, they support reliability evaluation through uncertainty metrics such as expected calibration error (ECE) and overconfident error rate (OC@0.8). Finally, they can be used to study whether a model can avoid noisy or redundant context when multiple clinically related but answer-irrelevant guideline statements are available.

What the datasets contain? Each benchmark item contains a medical question, multiple answer options, a gold option label, and the corresponding topic-specific keypoint corpus. The questions are written to test clinically meaningful distinctions,

including diagnostic thresholds, disease activity, disease severity, imaging recommendations, treatment indications, and follow-up decisions. Distractor options are often medically plausible in a broad sense, but incorrect under the specific guideline context. This design makes the benchmark more suitable for evaluating retrieval-augmented medical QA than simple factual recall: a model must retrieve the correct guideline keypoints and avoid being distracted by related but non-decisive keypoints.

Keypoint corpus construction. The keypoint corpus is not produced by an additional learned model. It follows the public XML structure of each guideline or consensus document: a guideline statement serves as the keypoint, and its supporting section serves as the passage. This deterministic conversion does not use later manual pairing. In URGen, keypoints are used for ranking because they are compact and discriminative, while the associated passages are provided to the LLM only after adaptive depth selection. Table S2 summarizes the construction protocol.

Public release. We plan to publicly release the benchmark content, including the question files, answer labels, keypoint identifiers, keypoint texts, supporting passages, and evaluation scripts. Since the data are derived from public guidelines and expert-curated questions rather than patient records, the benchmark does not contain protected health information. The release is intended to make the retrieval and calibration evaluation reproducible, and to facilitate future work on guideline-grounded medical QA.

B Implementation Details and Computational Cost

Inference configuration. Table S3 reports the measured stage-wise runtime for the Llama-3.1-8B experiments. URGen is training-free, so it does not require parameter updates, auxiliary model training, or a learned retriever. Its computation comes from two LLM-facing stages: keypoint ranking and answer generation.

The shared 2.59 GPU-s ranking cost is followed

¹<https://etj.bioscientifica.com/view/journals/etj/11/6/ETJ-22-0189.xml>

²<https://etj.bioscientifica.com/view/journals/etj/7/4/ETJ490384.xml>

Table S1: Summary of the expert-curated ophthalmology QA benchmarks and keypoint corpora. We use *keypoints* as the unified term throughout the supplementary material. Each keypoint is a concise guideline statement paired with its supporting passage.

Dataset	Domain	#Questions	#Keypoints	Format	Source document
TED	Thyroid eye disease	101	35	MCQA	<i>Management of Thyroid Eye Disease by the American and European Thyroid Associations</i> (Burch et al., 2022)
GD	Graves’ disease	97	22	MCQA	<i>2018 European Thyroid Association Guideline for the Management of Graves’ Hyperthyroidism</i> (Kahaly et al., 2018)

Table S2: Keypoint corpus construction protocol. The same protocol is used for TED and GD.

Step	Operation	Purpose
Guideline parsing	Extract guideline statements, including recommendations, diagnostic criteria, treatment rules, and management statements, from the public XML structure.	Preserve concise statements suitable for keypoint-level ranking.
Passage alignment	Use the supporting XML section associated with each extracted statement.	Deterministically pair compact keypoints with fuller context without a learned extractor.
Question curation	Construct multiple-choice questions around guideline-specific distinctions, thresholds, treatment decisions, and follow-up scenarios.	Evaluate whether models can use the correct guideline keypoints rather than broad medical priors.
Expert verification	Review question-answer pairs and answer labels with ophthalmology experts.	Reduce ambiguity and ensure that the ground-truth answer is clinically meaningful.
Retrieval packaging	Store each item as $(k_i; t_i)$, where k_i is the keypoint and t_i is the supporting passage.	Support keypoint-level ranking followed by passage-level answer generation.

by answer cost that grows with retrieved context. URGen therefore costs 0.55 GPU-s/question more than Fixed-1 but 1.15 less than Fixed-5 under this setup.

Parallelization and comparison with conventional RAG. URGen follows the same high-level dependency structure as conventional RAG: first select relevant information, then answer the question conditioned on the selected information. The difference is that URGen evaluates multiple ranking temperatures and multiple answer-generation temperatures. These branches are independent after model loading. Therefore, when sufficient GPU resources or batching capacity are available, the critical path is still one keypoint-ranking batch followed by one answer-generation batch, similar to the retriever-then-reader structure of conventional RAG. In other words, URGen increases total batched computation, but it does not introduce a long sequential reasoning chain.

This property gives URGen a practical advantage over iterative adaptive methods such as RQ-RAG (Chan et al., 2024). RQ-RAG refines or decomposes the query into multiple retrieval queries before answer generation, so later retrieval steps can depend on preceding generated queries. Such

query-refinement pipelines are harder to fully parallelize and can accumulate errors when intermediate rewritten queries drift from the clinical intent of the input question. URGen instead keeps the input medical question fixed and estimates retrieval ambiguity through multi-temperature keypoint ranking. This makes the adaptive step simple, parallelizable, and directly tied to the uncertainty of the keypoint distribution.

Reproducibility notes. All methods use the same question set, answer parser, and keypoint corpus unless otherwise stated. For keypoint ranking, we parse the ordered keypoint IDs from the JSON field `ordered_keypoints`. Each temperature contributes exactly one structured generation; malformed outputs are not regenerated. Regenerating until a ranking is parseable would selectively retain more coherent high-temperature samples and alter the rank-2 distribution, to which Fixed-2 is directly exposed. The one-attempt rule is therefore applied uniformly to all methods. Selected keypoint-passage blocks are concatenated in descending relevance order. For answer generation, we parse only the option label stored in the JSON field `answer`; rationales are retained for qualitative inspection but are not used when computing accu-

Table S3: Stage-wise inference runtime. All methods use Llama-3.1-8B in BF16 on one NVIDIA RTX 4090 (24GB) and the same temperature set. GPU-seconds are measured per question and exclude model loading.

Method	Retriever (GPU-s)	Answer (GPU-s)	Total (GPU-s/question)
<i>Fixed-1</i> RAG	2.59	1.19	3.78
<i>Fixed-5</i> RAG	2.59	2.89	5.48
URGen	2.59	1.74	4.33

Table S4: Serving-level comparison between conventional RAG, RQ-RAG, and URGen. URGen uses more batched samples than conventional RAG, but does not add a long sequential dependency chain.

Method	Main dependency	Parallelization	Adaptive signal	Practical implication
Conventional RAG	Retrieve once, then answer once.	Retrieval and answer stages are sequential, but each stage is simple.	Fixed retrieval depth or retriever score.	Efficient, but sensitive to the chosen retrieval depth.
RQ-RAG	Generate/refine multiple retrieval queries, retrieve, then answer.	Query-refinement steps may introduce sequential dependencies.	Query rewriting and retrieval expansion.	Flexible, but can be slower and can drift under noisy query refinement.
URGen	Rank keypoints under multiple temperatures, select depth, then answer under multiple temperatures.	Ranking samples and answer samples are independent and can be batched.	Entropy of keypoint ranking and answer distribution.	Training-free adaptive retrieval with a short critical path under sufficient resources.

racy, OC@0.8, or ECE.

C Additional Reliability and Sensitivity Analyses

Paired bootstrap comparisons. Table S5 gives the complete 10,000-sample paired question-level bootstrap results corresponding to the five-seed comparison in Table 3. The accuracy interval includes zero for Fixed-1 but excludes zero for Fixed-2 and Fixed-5. OC@0.8 differences exclude zero for every fixed depth, whereas the ECE interval excludes zero only for Fixed-5.

Temperature-count sensitivity. Table S6 jointly varies the numbers of retrieval and answer temperatures. Five temperatures give the highest mean accuracy and lowest OC@0.8; seven temperatures give the lowest ECE but require two additional samples at each stage. This comparison motivates the five-temperature setting without claiming that one setting dominates every metric.

Depth mapping. For the linear alternative, we set $h = u_{\text{context}} / \log K$ and $D_{\text{linear}} = \text{clip}_{[1,5]}(\lfloor 1 + 4h \rfloor)$. Table S7 shows that the exponential mapping has higher mean accuracy and lower OC@0.8, with similar ECE under the evaluated protocol.

Pooled candidate gallery. Table S8 evaluates TED questions after adding the 22 GD keypoints

to the 35-keypoint TED gallery. All three methods decline, but URGen remains strongest in the 57-keypoint pool: 4.55 percentage points above Fixed-1 and 8.12 above Fixed-5. This is a controlled disease-specific distractor test, not a web-scale retrieval experiment.

Normalized uncertainty ranges. Only the normalized entropies in Eq. 9 enter the fusion rule. Table S9 records their observed ranges and confirms that the fusion inputs stay within $[0, 1]$ on both datasets.

Low-context-uncertainty errors. Table S10 counts wrong answers among questions with the lowest context uncertainty. The nonzero counts expose the anticipated failure mode: a concentrated ranking can still be wrong. Because the benchmark has no gold keypoint labels, these counts characterize answer errors conditional on context uncertainty rather than directly validating the ranking itself.

D Retrieval Baseline Comparisons

We compare URGen with two groups of retrieval baselines. The first group uses dense embedding retrieval, including CLIP and BioMedCLIP, to test whether replacing the retriever alone is sufficient. The second group uses LLM-based ranking or query refinement, including a single-pass Llama ranker and RQ-RAG, to test whether

Table S5: Paired question-level bootstrap comparisons on TED (10,000 samples). Differences are URGen minus the named fixed-depth baseline; negative OC@0.8 and ECE differences favor URGen.

Comparison	Metric	Δ (pp)	95% CI (pp)	p
URGen – Fixed-1	Accuracy	+5.74	[−0.20, 11.68]	0.0630
	OC@0.8	−9.90	[−14.26, −5.54]	0.0002
	ECE	−8.00	[−13.59, 0.54]	0.0690
URGen – Fixed-2	Accuracy	+7.33	[0.79, 14.06]	0.0318
	OC@0.8	−7.33	[−11.49, −3.17]	0.0016
	ECE	−6.46	[−12.15, 2.27]	0.1766
URGen – Fixed-5	Accuracy	+10.69	[3.17, 18.42]	0.0052
	OC@0.8	−11.68	[−17.62, −5.94]	0.0002
	ECE	−12.13	[−18.89, −2.50]	0.0110

Table S6: Joint sensitivity to retrieval and answer temperature counts on TED with Llama-3.1-8B. Values are mean $\pm$ standard deviation over five seeds.

M_1	M_2	Temperatures	Acc. (%) $\uparrow$	OC@0.8 (%) $\downarrow$	ECE (%) $\downarrow$	Avg. depth
3	3	{0.3, 0.5, 0.7}	78.02 $\pm$ 1.29	7.52 $\pm$ 1.13	5.45 $\pm$ 1.03	3.06
5	5	{0.3, 0.4, 0.5, 0.6, 0.7}	80.40 $\pm$ 1.29	1.98 $\pm$ 0.00	6.80 $\pm$ 1.13	3.25
7	7	{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8}	79.60 $\pm$ 1.33	3.96 $\pm$ 0.70	4.00 $\pm$ 1.04	3.25

Table S7: Linear versus exponential uncertainty-to-depth mapping on TED with Llama-3.1-8B (five seeds).

Mapping	Acc. (%) $\uparrow$	OC@0.8 (%) $\downarrow$	ECE (%) $\downarrow$
Linear	78.81 $\pm$ 2.85	7.39 $\pm$ 1.63	6.60 $\pm$ 0.57
Exponential	80.40 $\pm$ 1.29	1.98 $\pm$ 0.00	6.80 $\pm$ 1.13

Table S8: TED accuracy with the TED-only 35-keypoint gallery and the pooled 57-keypoint TED+GD gallery.

Method	TED-35	Pooled-57	Change (pp)
Fixed-1	74.65	72.48	−2.18
Fixed-5	69.70	68.91	−0.79
URGen	80.40	77.03	−3.37

stronger LLM-based retrieval control can close the gap to URGen.

D.1 Dense Retriever Baselines

Dense retrievers map the question and candidate keypoints into an embedding space and select the nearest keypoint. This setting is efficient and simple, but it does not provide an explicit estimate of retrieval ambiguity. In our setting, dense retrieval also faces a mismatch: the keypoints are short clinical recommendations, while many questions require distinguishing between several semantically related guideline statements. Table S11 shows that BioMedCLIP improves over generic CLIP, but both remain below URGen.

Table S9: Observed ranges of normalized context and answer uncertainty used in fusion.

Dataset	$\tilde{u}_{\text{context}}$	$\tilde{u}_{\text{answer}}$
TED	[0.545, 0.995]	[0.000, 0.875]
GD	[0.581, 0.997]	[0.000, 0.964]

Table S10: Wrong answers in low-context-uncertainty subsets on TED.

Subset	Wrong	Subset size
Lowest 25%	2	26
Lowest 50%	5	51

The dense retriever results indicate that simply replacing the retrieval backbone is not enough. BioMedCLIP is better aligned with biomedical terminology than CLIP, but it still retrieves a fixed number of keypoints and cannot express whether the relevance distribution is sharp or diffuse. URGen uses the ranking distribution itself as a signal, allowing it to select fewer keypoints when one or two guideline statements are decisive, and more keypoints when the question requires broader clinical context.

D.2 LLM Ranking and RQ-RAG Baselines

We further compare against LLM-based ranking and query-refinement baselines in Table S12. The single-pass Llama ranker uses the same Llama-3.1-8B backbone to rank keypoints once, then retrieves only the top-1 keypoint. RQ-RAG refines

Table S11: Dense retrieval baselines on TED using Llama-3.1-8B. These methods retrieve a top-1 keypoint-passage block and do not provide calibrated context uncertainty.

Method	Retrieval source / procedure	Adaptive depth	Context uncertainty	Acc. (%) [†]
CLIP fixed-1 (Radford et al., 2021)	Generic CLIP-style embedding retrieval over the same keypoint corpus; top-1 keypoint only.	×	×	69.31
BioMedCLIP fixed-1 (Zhang et al., 2025)	Biomedical vision-language embedding retrieval over the same keypoint corpus; top-1 keypoint only.	×	×	71.29
URGen	Multi-temperature LLM keypoint ranking with entropy-guided adaptive depth.	✓	✓	80.40

Table S12: Comparison with LLM-based ranking and RQ-RAG on TED using Llama-3.1-8B. RQ-RAG improves retrieval flexibility through query refinement, while URGen estimates keypoint-ranking uncertainty and adaptively selects retrieval depth.

Method	Retrieval procedure	Adaptive depth	Context unc.	Answer unc.	Acc. (%) [†]	OC@0.8 (%) [↓]	ECE (%) [↓]
Single-pass Llama ranker	Rank keypoints once with Llama-3.1-8B, then use only the top-1 keypoint-passage block.	×	×	✓	71.29	12.87	15.36
RQ-RAG (Chan et al., 2024)	Refine/decompose the question into up to five retrieval queries before answer generation.	✓	×	✓	71.29	5.94	17.31
URGen	Multi-temperature keypoint ranking, entropy-guided adaptive depth, and fused context-answer uncertainty.	✓	✓	✓	80.40	1.98	6.80

or decomposes the input into up to five retrieval queries before answer generation. Although RQ-RAG introduces adaptive query reformulation, it does not explicitly estimate context uncertainty from the ranking distribution. URGen differs by keeping the input clinical question fixed and using multi-temperature ranking to infer how many keypoints should be retained.

The gap between RQ-RAG and URGen suggests that adaptive query expansion alone is not sufficient for this benchmark. In medical guideline QA, query refinement can generate related search queries, but it may still retrieve redundant or clinically adjacent keypoints that do not decide the answer. URGen instead estimates whether the ranked keypoint distribution is concentrated or diffuse. This allows the model to contract the context when the decisive keypoints are clear and expand the context when the evidence is ambiguous.

E Uncertainty Analysis

Relation to semantic graph uncertainty. Free-form uncertainty metrics such as Deg, Ecc, EigV, and word-sequence entropy (Lin et al., 2024) estimate uncertainty from the lexical or semantic di-

versity of multiple generated answers. They are most informative when generations are long and can form multiple semantic clusters. Our task is a constrained multiple-choice setting: each generation contains a discrete option choice and a short rationale. Under this format, Jaccard- or NLI-based similarity graphs are dominated by whether two outputs select the same option, so these graph-based metrics largely reduce to functions of the option vote distribution.

Why answer entropy alone is insufficient?

The main-paper ablation already isolates this effect: replacing fused uncertainty with answer entropy only leaves the final predictions unchanged but worsens OC@0.8 from 1.98% to 6.93% and ECE from 6.80% to 12.31%. This shows that disagreement among generated options is not sufficient for reliable calibration; context ambiguity must also be modeled. In clinical QA, a model can repeatedly emit the same answer while relying on an incomplete or mismatched keypoint set. URGen therefore combines answer uncertainty with context uncertainty derived from the keypoint ranking distribution.

Interpretation of context uncertainty. Context uncertainty should not be interpreted as a direct measure of disease uncertainty. It measures instability in the retrieved keypoint distribution. A high context-uncertainty score indicates that multiple keypoints are similarly plausible for the question, which may reflect broad clinical scope, missing decisive evidence, or ambiguous wording. A low context-uncertainty score indicates that the ranking is concentrated on a small number of keypoints. Fusing this score with answer uncertainty makes the final uncertainty sensitive to both retrieval ambiguity and answer-generation ambiguity.

F More Case Studies

Table S13 provides an additional TED case study. The example shows that fixed-depth retrieval may include weakly related keypoints even when it eventually predicts the correct option. URGen selects a smaller keypoint set and produces a substantially lower uncertainty score. This complements the quantitative results by illustrating how adaptive depth suppresses redundant context while preserving the decisive keypoints.

Case analysis. The vanilla model chooses contrast-enhanced CT despite the guideline-specific preference for noncontrast CT in this surgical-planning context. Fixed-5 RAG corrects the answer, but its uncertainty remains high because the retrieved keypoints include several broad TED assessment and referral statements that are only weakly related to the imaging modality. URGen retrieves a smaller set of high-ranking key-

points and assigns a lower uncertainty score to the same correct answer. This example illustrates an important distinction: retrieving more keypoints can increase topical coverage, but it can also introduce redundant or weakly relevant context that makes the model less certain. URGen’s adaptive depth selection helps balance evidence sufficiency and context compactness.

Qualitative takeaway. The case study supports the quantitative trend in the main paper. When the keypoint distribution is concentrated, URGen contracts the context and avoids unnecessary guideline material. When the keypoint distribution is diffuse, the method can expand the context to reduce the risk of missing key clinical information. This behavior is especially important for guideline-grounded medical QA, where several keypoints may be semantically close but only one or two are clinically decisive for the final option.

G Prompt Templates

URGen is implemented in a training-free manner using two prompt templates. The first prompt ranks candidate guideline keypoints for retrieval, and the second prompt answers the multiple-choice question using only the selected keypoint-passage blocks. The prompts enforce structured JSON output to make retrieval ranking and answer extraction reproducible.

Table S13: Additional TED multiple-choice case study using Llama-3.1-8B. The example compares vanilla prompting, fixed-5 RAG, and URGen.

Case Study 2	
Model / Corpus	Llama-3.1-8B / TED
Question & Choices	For TED patients preparing for surgery, the preferred imaging modality is: A. Contrast-enhanced CT B. Noncontrast CT C. Ultrasound D. Nuclear medicine scan
GT	B. Noncontrast CT
Vanilla	Answer: A. Contrast-enhanced CT (Unc.: 5.70%)
Fixed-5 RAG	URGen
Answer: B. Noncontrast CT (Unc.: 36.64%)	Answer: B. Noncontrast CT (Unc.: 9.21%)
Selected 5 Keypoints	Selected 2 Keypoints
1. 2.1 Activity and severity definitions for patients with TED, CAS score definition.	1. 2.1 Activity and severity definitions for patients with TED, CAS score definition. (Ranking Score: 0.49)
2. 4.1.2 Assessment of TED should include activity, severity, trend over time, and impact on daily living.	2. 4.1.2 Assessment of TED should include activity, severity, trend over time, and impact on daily living. (Ranking Score: 0.14)
3. 4.1.1 Endocrinologists should be familiar with TED examination basics.	
4. 3.3 Ophthalmologists should be involved in moderate-to-severe and sight-threatening TED management.	
5. 7.3.1.1 TED surgery should be performed by an experienced orbital surgeon.	

PROMPT TEMPLATE: MULTI-TEMPERATURE KNOWLEDGE RETRIEVAL

SYSTEM / ROLE

You are a clinical guideline assistant. Your job is to arrange candidate guideline keypoints by how useful they are for answering the question.

INPUT

Question: <QUESTION>

Options: (A) <A> (B) (C) <C> (D) <D> (E) <E>

Candidate keypoints (ID and text):

[k_1] <KEYPOINT_TEXT_1>

[k_2] <KEYPOINT_TEXT_2>

...

[k_N] <KEYPOINT_TEXT_N>

Temperature tag: <TEMP_TAG> (logging only)

INSTRUCTIONS

- Return a **single ordered list** of keypoint IDs, from most useful to least useful.
- Rank by how well each keypoint helps distinguish among options, prioritizing diagnostic criteria, thresholds, contraindications, and treatment rules.
- Output **strict JSON only** following the schema below, with no extra text.

HARD RULE

Do not answer the question.

OUTPUT JSON SCHEMA

```
{
  "version": "urgen-step1-v1",
  "tau": "<TEMP_TAG>",
  "ordered_keypoints": ["k_7", "k_3", "k_12", "...", "k_N"]
}
```

Figure S1: Prompt template used in multi-temperature knowledge retrieval.

PROMPT TEMPLATE: MULTI-ANSWER AGGREGATION

SYSTEM / ROLE

You are a cautious clinical QA assistant. Answer the multiple-choice question using only the retrieved keypoint-passage blocks below.

INPUT

Question: <QUESTION>

Options: (A) <A> (B) (C) <C> (D) <D> (E) <E>

Retrieved keypoint-passage blocks (selected by ZARD, *D* blocks):

[k_1] Keypoint: <KP_1> Passage: <P_1>

[k_2] Keypoint: <KP_2> Passage: <P_2>

... (repeat until [k_D])

Temperature tag: <TEMP_TAG> (logging only)

INSTRUCTIONS

- Choose exactly one option from {A,B,C,D,E}.
- Use only keypoint-passage blocks [k_1]–[k_D]. Do not use outside medical knowledge.
- Provide a short justification in 1–2 sentences and cite keypoint IDs, e.g., [k_2], [k_5].
- If evidence is insufficient or conflicting, set `insufficient_evidence=true`.
- Output strict JSON only with the schema below.

OUTPUT JSON SCHEMA

```
{
  "version": "urgen-step2-v1",
  "tau": "<TEMP_TAG>",
  "answer": "C",
  "rationale": "...",
  "evidence_used": ["k_2", "k_5"],
  "insufficient_evidence": false
}
```

Figure S2: Prompt template used in multi-answer aggregation.