CURA: Calibrated Uncertainty with Retrieval and Agents for Trustworthy Multimodal Medical Decision Support

Ruichen Zhang[1,*], Jessie Dong[2,*], Jiayi Xin[2], Jenna Ballard[2], Cameron Beeche[2], Rakesh Sharma[2], Xinyu Zhao[1], Nicholas Konz[1], Qi Long[2], Walter Witschey[2], and Tianlong Chen[1]

[1] University of North Carolina at Chapel Hill, United States
tianlong@cs.unc.edu
[2] University of Pennsylvania, United States
*Equal contribution.

Abstract. Large language models (LLMs) for medicine frequently fail in safety-critical settings due to miscalibrated confidence, brittle deliberation, and uncertainty signals that are hard to act on. We present CURA, a reliability-first multimodal medical decision-support framework that couples knowledge-graph grounded retrieval with a heterogeneous clinical council, fuses opinions via Bayesian belief aggregation, and exposes a single calibrated posterior to downstream decision policies. CURA instantiates two reliability overlays: selective abstention (withhold answers when uncertain) and split conformal prediction sets (set-valued predictions with calibrated marginal coverage). CURA attains strong accuracy on four medical QA benchmarks: 88.5% (MedQA), 79.5% (MedMCQA), 89.3% (MMLU-Medical), and 97.5% (QA4MRE). Selective abstention yields a clear risk–coverage trade-off (e.g., MedMCQA: risk is 20.8% at 80% coverage; AURC=0.213; Fig. 3). Conformal sets meet target coverage (MedMCQA α=0.05: 99.2% [97.1%, 99.9%]) and rescue most top-1 errors (97.0%). Interactive AgentClinic evaluations show that uncertainty-driven test acquisition improves diagnosis on complex cases (+15pp on NEJM). Finally, across adaptive-council configurations, we find that naïvely scaling experts can degrade both accuracy and calibration; the entropy threshold—not posterior max-probability—is the primary early-stopping control. CURA provides calibrated answers with uncertainty and cost-aware escalation. Code is available at `https://github.com/UNITES-Lab/CURA`.

Keywords: Medical decision support · Calibrated uncertainty · Conformal prediction · Multi-agent deliberation · Knowledge graphs.

1 Introduction

Clinical decision-making is inherently deliberative: complex cases are discussed in multidisciplinary team meetings where disagreement often signals uncertainty more reliably than any single opinion [17]. LLMs promise broad medical

knowledge [25,21,26], but current systems often behave like monolithic oracles—producing a single answer with unreliable confidence [9,11]. This gap is particularly consequential in medicine: *overconfident* errors can mislead triage and referral, while *underconfident* correct answers can trigger unnecessary testing [2,20].

We present CURA, a reliability-first multimodal medical decision support framework. CURA integrates three complementary ideas: **(i) knowledge grounding** via a bidirectional LLM↔KG pipeline with probabilistic graph exploration [7,15,8], **(ii) cognitive diversity** via a heterogeneous clinical council whose opinions are fused with Bayesian belief aggregation [6,24], and **(iii) uncertainty-to-action overlays** that expose calibrated uncertainty to downstream decision policies. Rather than assuming "more deliberation is always better," we explicitly study *when* deliberation helps and *how* to allocate it under cost constraints. Our adaptive-council analysis yields a key practical insight: blind scaling of experts can hurt on well-specified multiple-choice benchmarks, motivating entropy-based gating and conformal prediction sets as safety overlays.

Contributions. (i) **Framework:** CURA combines co-evolutionary LLM↔KG grounding with Monte Carlo KG exploration, a heterogeneous clinical council, and Bayesian belief aggregation. (ii) **Reliability overlays:** we operationalize uncertainty via selective abstention with risk–coverage evaluation, split conformal prediction sets (set-valued outputs with calibrated marginal coverage), and an uncertainty-driven commit/continue policy for AgentClinic-style interactive diagnosis. (iii) **Cost–calibration insight:** across adaptive-council configurations, we show that single-expert inference is Pareto-optimal, entropy is the dominant early-stopping knob for $\tau_p \leq 0.8$, and overly strict τ_p (e.g., 0.9) becomes a binding constraint that forces escalation.

Related Work. Retrieval-augmented generation [15] and structured medical knowledge can improve factuality in clinical QA [30,29], but safety remains limited by hallucinations and miscalibrated confidence. Selective prediction trades coverage for lower conditional risk, while split conformal prediction provides set-valued outputs with finite-sample marginal coverage control, offering a practical safety net when single best guesses may be wrong. Multi-agent deliberation can diversify reasoning through debate [6] and council-style aggregation [4], but may also introduce noise or anchoring effects, motivating cost-aware escalation policies that allocate additional experts only when uncertainty warrants it [9,4]. However, existing frameworks lack a unified mechanism that couples calibrated uncertainty with both abstention and set-valued prediction overlays for medical decision support.

2 Method

Fig. 1 overviews CURA. We describe core components concisely and emphasize the reliability overlays evaluated in this paper.

Co-evolutionary LLM↔KG Grounding. CURA maintains multiple domain knowledge graphs (clinical, pharmacology, imaging, neuro-medical) constructed

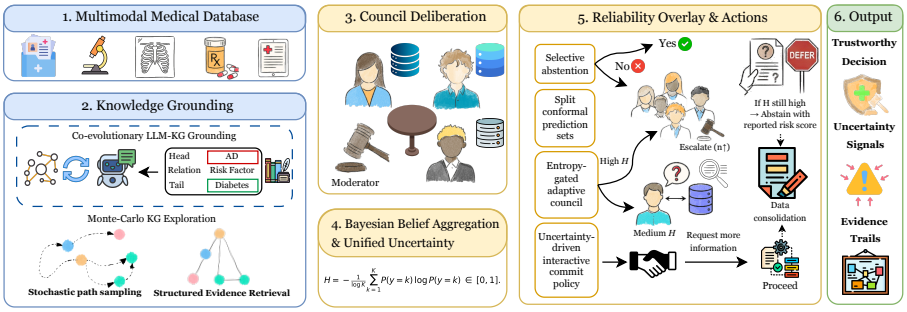

Fig. 1. CURA overview. KG-grounded retrieval, heterogeneous clinical council, Bayesian aggregation, and reliability overlays (selective abstention, conformal prediction sets, entropy-gated escalation, uncertainty-driven interaction).

from literature and curated sources. At inference, we (1) extract biomedical entities, (2) match entities to KG nodes via embedding similarity [18,27], and (3) retrieve evidence via a hybrid of neighborhood expansion and *Monte Carlo path sampling* [3,32]. Unlike deterministic shortest-path retrieval, stochastic traversal surfaces diverse multi-hop chains, which is valuable for atypical or rare presentations. Retrieved triples are reranked and injected as structured context to ground the LLM response. To reduce evaluation leakage, we exclude benchmark datasets and their official explanations/answer keys from the KG and retrieval store; retrieved evidence is drawn from general medical references rather than dataset artifacts.

Implementation details. For Monte Carlo KG exploration, we sample multiple multi-hop paths per query (with a fixed hop limit) and score candidate triples by a weighted combination of entity-match confidence and path relevance to the query; top-ranked evidence is injected as structured context. Expert personas use a fixed prompt template and report both an option and a self-reported confidence in $(0, 1)$. Unless stated otherwise, we set expert weights uniformly (w_i=1) and aggregate in log space as described below.

Heterogeneous Clinical Council. We use a multidisciplinary council of specialist personas (e.g., attending, internist, pharmacologist, diagnostician, safety officer). Each expert independently reviews the query, evidence, and available images, then reports an answer choice with calibrated confidence. Experts operate with isolated, non-communicating knowledge bases and user memories; the safety officer additionally flags contraindications and high-risk recommendations.

Bayesian Belief Aggregation. Given K answer options and n experts, each expert i outputs an option o_i and confidence $c_i \in (0, 1)$ with weight w_i. We first compute unnormalized log scores $\ell_k = \log(1/K) + \sum_i w_i \log P(o_i \mid y{=}k)$, where $P(o_i \mid y{=}k) = c_i$ if $o_i = k$ and $P(o_i \mid y{=}k) = (1{-}c_i)/(K{-}1)$ otherwise. We then normalize across answer options as $P(y = k \mid o_{1:n}) = \exp(\ell_k)/\sum_{k'=1}^{K} \exp(\ell_{k'})$. Uncertainty is normalized Shannon entropy $H = -(\log K)^{-1} \sum_{k=1}^{K} P(y = k \mid o_{1:n}) \log P(y{=}k \mid o_{1:n}) \in [0, 1]$. This posterior and entropy serve as the unified

Table 1. Uncertainty-to-action overlays in CURA. A single aggregated posterior enables abstention, set-valued prediction, cost-aware escalation, and interactive commit policies.

Overlay	Trigger (from posterior)	Output / action
Selective abstention	Answer iff $H(x) \leq \tau_H$	Withhold answer; defer or interact
Conformal sets	Split conformal at miscoverage α	Return set $S(x)$ with target coverage
Adaptive council	Stop iff $p_{\max} \geq \tau_p \wedge H \leq \tau_H$	Allocate more experts only when needed
AgentClinic commit	Commit iff $p_{\max} \geq \tau_p \wedge H \leq \tau_H$	Diagnose vs. acquire more information

uncertainty interface for adaptive deliberation, conformal sets, and interactive policies [1,14].

Selective Abstention (Risk–Coverage). Given the aggregated posterior $p(\cdot|x)$, we quantify uncertainty with normalized entropy $H(x) = -\sum_k p_k \log p_k / \log K$. Selective prediction answers only when $H(x) \leq \tau_H$ (equivalently, abstains when uncertain). Coverage is the fraction of examples answered; risk is the error rate conditional on answering. We report the full risk–coverage curve by sweeping τ_H (200 points) and summarize performance via AURC (area under the risk–coverage curve; lower is better).

Split Conformal Prediction Sets. Given posterior probabilities $p_k = P(y = k|x)$, we compute nonconformity scores on a calibration set: $s_i = 1 - p_{y_i}$. For target miscoverage α, we set threshold $\hat{q}$ as the conservative split-conformal quantile $\hat{q} = \text{Quantile}_{\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil / n_{\text{cal}}}(s)$ (using `higher` quantile interpolation). The prediction set is $S(x) = \{k : p_k \geq 1 - \hat{q}\}$; if $S(x)$ is empty, we fall back to argmax. We report empirical coverage with Clopper–Pearson confidence intervals.

Uncertainty-Driven Commit Policy (AgentClinic). In interactive diagnosis, the agent iteratively asks questions or requests tests until it commits. We commit when either (i) the agent explicitly chooses a `DIAGNOSE` action, or (ii) its posterior satisfies $p_{\max} \geq \tau_p$ and $H \leq \tau_H$.

Entropy-Gated Adaptive Council. We allocate expert deliberation as a limited resource. Starting from a small council (stage sizes $[2, 3, 5, 7]$), we add new experts only when the current posterior is insufficiently confident: stop if $p_{\max} \geq \tau_p \wedge H \leq \tau_H$; else escalate to the next stage. Earlier experts are never re-invoked; the moderator re-synthesizes using all collected opinions.

3 Experiments

Setup. We evaluate across three settings that reflect realistic clinical usage: (i) multiple-choice medical QA (MedQA [10], MedMCQA [22], MMLU-Medical [31], QA4MRE [23]), (ii) visual medical QA (VQA-RAD [13]), and (iii) interactive diagnosis (AgentClinic [24] on MedQA and NEJM scenarios).

Table 2. Main QA benchmark results. Accuracy (%) on four medical QA datasets. Best in **bold**, second-best underlined.

	MedQA	MedMCQA	MMLU	QA4MRE	AVG
Biomedical LLMs					
Biomistral-7B [12]	44.7	49.5	53.1	68.6	54.0
Meditron-70B [5]	50.0	44.8	79.6	51.4	56.4
ClinicalCamel-70B [28]	50.0	64.3	83.7	68.6	66.7
Retrieval-augmented LLMs					
Clinfo.ai [19]	54.3	77.0	81.3	67.7	70.1
DALK [16]	57.9	75.2	85.4	71.4	72.6
GPT-4.1	60.9	68.1	80.1	**100.0**	77.3
CURA	**88.5**	**79.5**	**89.3**	97.5	**88.7**

Unless stated otherwise, we use GPT-4.1 (Azure deployment), temperature = 0, and fixed sampling seed 42. We distinguish two single-model baselines used for different purposes. In Table 2, "GPT-4.1" denotes a plain single-shot baseline without CURA's KG grounding, specialist prompting, or Bayesian aggregation. In Table 3 and reliability overlays (selective abstention and conformal sets below), "Single (1 expert)" denotes the CURA single-specialist variant with the same grounding and answer-extraction interface, but without multi-expert deliberation. ECE is computed as the expected calibration error by binning predictions by confidence (lower is better). Tokens denote the total number of input+output tokens consumed per example (including retrieved context and deliberation prompts) as reported by the API; AgentClinic cost is estimated in USD from token usage under the deployed endpoint pricing. Adaptive-council experiments evaluate a grid of entropy thresholds τ_H and posterior thresholds τ_p over a fixed set of stage sizes; we report representative operating points and aggregate trends over all evaluated configurations. Unless explicitly noted, comparisons isolate the component under study while keeping grounding and answer-extraction constant (e.g., Table 3 varies only deliberation size/control policy).

Benchmark Comparison. Table 2 summarizes results; CURA improves over most baselines under our evaluation protocol.

Adaptive Council: Cost–Accuracy–Calibration Trade-Offs. Table 3 reports representative operating points for single-expert, fixed-size councils, and adaptive councils under a shared token budget perspective. Single-expert inference is Pareto-optimal (tokens → accuracy) on both benchmarks, achieving 77.3% on MedMCQA (ECE=0.058; 1,387 tokens/sample) and 90.8% on MedQA (ECE=0.052; 2,025 tokens/sample). Fixed multi-expert panels substantially increase compute and can degrade both accuracy and ECE. Fig. 2a visualizes the tokens–accuracy frontier and highlights that, for well-specified multiple-choice QA, blind expert scaling is not a free gain. This motivates treating deliberation as a resource to allocate, rather than a default.

Table 3. Adaptive council operating points. Accuracy is reported with 95% Wilson CIs over questions. ⋆ = Pareto-optimal (tokens → accuracy).

Benchmark	Config	Acc. (%) [95% CI]	ECE ↓	Tokens	Experts
MEDMCQA	Adaptive (τ_p=0.8, τ_H=0.5)	76.2 [72.3, 79.7]	0.1095	5,266	2.7
MEDMCQA	Fixed-7	72.7 [68.7, 76.5]	0.1218	15,591	7.0
MEDMCQA	Single (1 expert) ⋆	77.3 [73.3, 80.7]	0.0584	1,387	1.0
MEDQA	Adaptive (τ_p=0.6, τ_H=0.5)	75.8 [71.9, 79.3]	0.0987	9,685	3.2
MEDQA	Adaptive (τ_p=0.7, τ_H=0.3)	76.2 [72.3, 79.7]	0.0722	25,858	6.7
MEDQA	Fixed-7	90.4 [87.5, 92.7]	0.1192	20,377	7.0
MEDQA	Single (1 expert) ⋆	90.8 [87.9, 93.0]	0.0519	2,025	1.0

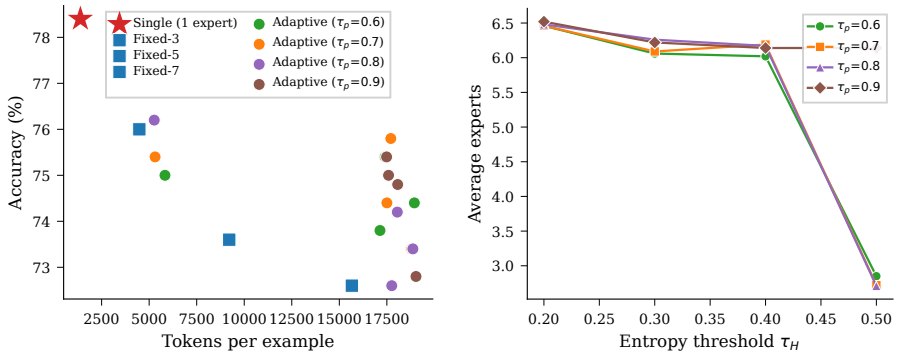

Fig. 2. Cost-aware deliberation. **(a)** Tokens vs. accuracy Pareto scatter on MedMCQA for single, fixed-size councils, and adaptive policies. **(b)** Entropy gating dominates escalation: average experts as a function of entropy threshold τ_H for several posterior thresholds τ_p.

Mechanistic Insight: What Controls Early Stopping? We analyze escalation behavior as a function of the entropy threshold τ_H and posterior threshold τ_p. For $\tau_p \leq 0.8$, τ_H is the binding constraint: at τ_H=0.5, varying τ_p (0.6/0.7/0.8) yields nearly identical average experts (≈2.7–2.8) and full-escalation rate (≈11–12%). In contrast, τ_p=0.9 becomes a case: posterior maxima rarely exceed 0.9, so τ_p forces escalation even when entropy gating is relaxed (behavior comparable to strict $\tau_H \leq 0.4$). Fig. 2b summarizes this regime change and explains why entropy is the dominant early-stopping knob in typical posterior ranges.

Selective Abstention: Risk–Coverage Trade-Offs.

We evaluate selective abstention on fixed 500-question subsets of MedQA and MedMCQA by ranking questions by normalized entropy H and answering only the least-uncertain fraction, producing a risk–coverage curve (risk = error rate on answered questions; coverage = fraction answered). Fig. 3a shows consistent risk reduction as coverage decreases on both datasets. At moderate coverage, selective abstention lowers conditional error on both benchmarks: on MedMCQA, risk is 20.8% at 80% coverage; on MedQA, risk is 6.8% at 80% coverage. AURC values

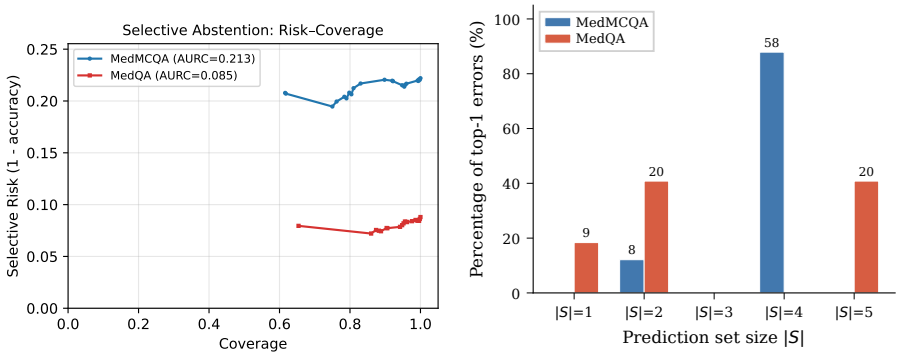

Fig. 3. Reliability overlays convert calibrated uncertainty into actionable decisions. **(a)** Selective abstention: risk–coverage curves on MedQA and MedMCQA obtained by thresholding normalized entropy. **(b)** Conformal safety net: distribution of conformal set sizes conditioned on the top-1 prediction being wrong; rescue at α=0.05 is 97.0% (MedMCQA) and 73.5% (MedQA). Numbers above bars denote counts.

are 0.213 (MedMCQA) and 0.085 (MedQA). These results substantiate selective abstention as a practical reliability overlay: instead of emitting overconfident answers, the system can abstain and defer to a human or trigger interactive information gathering (AgentClinic).

Conformal Prediction Sets: Coverage and Clinical Rescue. We evaluate split conformal prediction on MedMCQA and MedQA using a seeded permutation split into calibration and test subsets. Table 4 reports empirical coverage with 95% Clopper–Pearson CIs and clinical safety metrics. At α=0.05, conformal sets achieve 99.2% coverage on MedMCQA and 94.8% on MedQA. Clinically, conformal sets provide a safety net: conditional on the model's top-1 prediction being wrong, the ground truth remains in the set in 97.0% (MedMCQA) and 73.5% (MedQA) of cases. Fig. 3b visualizes the set-size distribution on the top-1-wrong subset; conformal misses correspond to high-risk failure modes—often singleton sets containing an incorrect answer—highlighting where additional safeguards are needed. Across evaluated α levels, misses occur only when the underlying top-1 is already wrong (i.e., within the evaluated subsets, we do not observe cases where the argmax is correct but the conformal set excludes the ground truth).

Interactive Evaluation (AgentClinic). We evaluate AgentClinic-style interaction under three policies: *immediate* diagnosis, *fixed* 5 turns, and *uncertainty-driven* (commit when $p_{\max} \geq 0.85$ and $H \leq 0.3$). On MedQA scenarios, uncertainty-driven interaction improves accuracy from 14.0% (immediate) to 31.0% (+17pp), while increasing cost. On NEJM scenarios, uncertainty-driven interaction improves accuracy from 0.0% to 15.0%. These results support a practical interpretation: interaction helps primarily when the initial information is insufficient, but should be triggered selectively due to cost.

Table 4. Split conformal prediction sets using a seeded permutation split into calibration and test subsets. Coverage is reported with 95% Clopper–Pearson CIs. Rescue is the fraction of top-1 errors where the correct answer remains in the set. Singleton is the fraction of test examples with $|S|$=1.

Benchmark	α	Coverage (95% CI)	Avg $\|S\|$	Rescue ↑	Singleton	Missed
MedMCQA	0.05	99.2 [97.1, 99.9]	3.68	97.0%	5.2%	2/250
MedMCQA	0.10	94.0 [90.3, 96.6]	3.56	77.3%	11.6%	15/250
MedMCQA	0.20	82.4 [77.1, 86.9]	1.74	33.3%	69.6%	44/250
MedQA	0.05	94.8 [91.3, 97.2]	3.96	73.5%	11.2%	13/250
MedQA	0.10	91.6 [87.5, 94.7]	3.92	57.1%	15.6%	21/250
MedQA	0.20	82.0 [76.7, 86.6]	1.10	8.2%	92.8%	45/250

Table 5. AgentClinic results. Exact-match accuracy with 95% Clopper–Pearson confidence intervals, average turns, and cost.

Dataset	Mode	n	Accuracy	95% CI	Avg Turns	Cost ($)
NEJM	Immediate	100	0.0%	[0.00, 0.04]	1.0	0.10
	Fixed (T=5)	100	12.0%	[0.06, 0.20]	4.0	0.94
	Uncertainty	100	15.0%	[0.09, 0.24]	6.3	1.93
MedQA	Immediate	100	14.0%	[0.08, 0.22]	1.0	0.15
	Fixed (T=5)	100	31.0%	[0.22, 0.41]	3.9	1.27
	Uncertainty	100	31.0%	[0.22, 0.41]	5.6	2.92

Table 6. VQA-RAD multimodal QA. Deliberation helps primarily on open-ended questions.

Mode	Accuracy	Closed Acc.	Open Acc.	Avg Latency (ms)
Baseline	82.0%	74.5%	91.1%	3,098
Expert panel	85.0%	74.5%	97.8%	12,383
Expert debate	84.0%	76.4%	93.3%	6,914

Visual Medical QA (VQA-RAD). To substantiate the multimodal setting, we evaluate CURA on VQA-RAD with image + question inputs. Table 6 compares a single-model baseline against two deliberative variants. The expert panel improves overall accuracy from 82.0% to 85.0%, with gains concentrated in open-ended questions (91.1%→97.8%), consistent with deliberation being most helpful when the answer space is under-specified.

4 Discussion and Conclusion

Our results highlight three practical lessons for deploying deliberative medical assistants. First, calibrated uncertainty can be operationalized: selective abstention reduces conditional risk, while conformal sets provide a set-valued safety net on top-1 failures (Fig. 3, Table 4). Second, more experts are not inherently better on well-specified multiple-choice QA: across adaptive-council configurations, single-

expert inference is Pareto-optimal and aggressive escalation can worsen both accuracy and calibration while consuming more tokens (Table 3, Fig. 2). Third, interaction helps when initial information is insufficient (AgentClinic), but should be triggered selectively due to cost. Overall, CURA unifies KG-grounded retrieval, Bayesian uncertainty, and reliability overlays for multimodal medical decision support, providing actionable uncertainty for safer, cost-aware deployment.

Acknowledgements

This work was partially funded by the National Institutes of Health (NIH) and Advanced Research Projects Agency for Health (ARPA-H) under grant numbers R01EB033387, 1R01EB037101-01, 5R01HL171709-02, 5R01HL169378-04, 1R21EB036734-01A1, 1OT2OD038048-01, R01EB036016, R01MH143267, and 2T32EB020087-11. The funder had no role in the study design, analysis, interpretation of the results, or preparation of the manuscript. We also acknowledge the creators and maintainers of the public datasets used in this study.

Disclosure of Interests

The authors have no competing interests to declare.

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification (2022)
2. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., et al.: Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis. Journal of Medical Internet research **26**(1), e53164 (2024)
3. Chen, J., Liu, G., He, S., Luo, K., Xu, Y., Zhao, J., Liu, K.: Search-in-context: Efficient multi-hop qa over long contexts via monte carlo tree search with dynamic kv retrieval. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 26443–26455 (2025)
4. Chen, X., Yi, H., You, M., Liu, W., et al.: Enhancing diagnostic capability with multi-agents conversational large language models. npj Digital Medicine **8**(1), 159 (2025). https://doi.org/10.1038/s41746-025-01550-0
5. Chen, Z., Cano, A.H., Romanou, A., et al.: Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079 (2023)
6. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023)
7. Edge, D., Trinh, H., et al.: From local to global: A graph RAG approach to query-focused summarization (2025). https://doi.org/10.48550/arXiv.2404.16130, https://arxiv.org/abs/2404.16130
8. Gao, G., Li, Z., Yuan, C., Li, J., Jianzhuo, W., Zhang, Y., Jin, X., Li, B., Hu, W.: D-RAG: Differentiable retrieval-augmented generation for knowledge graph question answering. In: Proceedings of the 2025 Conference on Empirical Methods in Natural

Language Processing. pp. 35398–35417. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.1793, https://aclanthology.org/2025.emnlp-main.1793/
9. Huang, Y., Liu, Y., Thirukovalluru, R., Cohan, A., Dhingra, B.: Calibrating long-form generations from large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 13441–13460 (2024)
10. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences (2021)
11. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., et al.: Language models (mostly) know what they know (2022). https://doi.org/10.48550/arXiv.2207.05221, https://arxiv.org/abs/2207.05221
12. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 5848–5864 (2024)
13. Lau, J.J., Gayen, S., et al.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data (2018)
14. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018). https://doi.org/10.1080/01621459.2017.1307116, https://doi.org/10.1080/01621459.2017.1307116
15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: NeurIPS (2020)
16. Li, D., Yang, S., Tan, Z., Baik, J.Y., Yun, S., Lee, J., Chacko, A., Hou, B., Duong-Tran, D., Ding, Y., et al.: Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 2187–2205 (2024)
17. List, H., Kristensen, D.B., Graumann, O.: "the highest decision-making level" – multidisciplinary team meetings as boundary spaces. Social Science & Medicine **371**, 117886 (2025). https://doi.org/10.1016/j.socscimed.2025.117886
18. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 4228–4238 (2021)
19. Lozano, A., Fleming, S.L., Chiang, C.C., Shah, N.: Clinfo.ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In: Pacific Symposium on Biocomputing 2024. vol. 29, pp. 8–23. World Scientific (2024). https://doi.org/10.1142/9789811286421_0002, https://doi.org/10.1142/9789811286421_0002
20. Manakul, P., Liusie, A., Gales, M.J.F.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models (2023)
21. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
22. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of the Conference on Health, Inference, and Learning. Proceedings of Machine Learning Research, vol. 174, pp. 248–260. PMLR (2022)
23. Peñas, A., Hovy, E.H., Forner, P., Rodrigo, Á., Sutcliffe, R.F.E., Morante, R.: Qa4mre 2011-2013: Overview of question answering for machine reading evaluation.

In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization (CLEF 2013). LNCS, vol. 8138, pp. 303–320. Springer (2013)
24. Schmidgall, S., Ziaei, R., Harris, C., et al.: Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments (2025)
25. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., et al.: Large language models encode clinical knowledge. Nature (2023)
26. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., et al.: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3), 943–950 (2025). https://doi.org/10.1038/s41591-024-03423-7, https://doi.org/10.1038/s41591-024-03423-7
27. Sung, M., Jeon, H., Lee, J., Kang, J.: Biomedical entity representations with synonym marginalization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 3641–3650 (2020)
28. Toma, A., Lawler, P.R., Ba, J., Krishnan, R.G., Rubin, B.B., Wang, B.: Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. arXiv preprint arXiv:2305.12031 (2023)
29. Vach, M., Gliem, M., Weiss, D., Ivan, V.L., Hauke, F., Boschenriedter, C., Rubbert, C., Caspers, J.: Evaluating retrieval augmented generation-enhanced large language models for question answering on german neurovascular guidelines. Clinical Neuroradiology **36**, 119–127 (2026). https://doi.org/10.1007/s00062-025-01562-z, published online 2 September 2025
30. Xia, P., Zhu, K., et al.: Rule: Reliable multimodal rag for factuality in medical vision language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1081–1093 (2024)
31. Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 6233–6251 (2024)
32. Xiong, G., Li, H., Zhao, W.: Mcts-kbqa: Monte carlo tree search for knowledge base question answering (2025)