Principled Context Engineering for RAG: Statistical Guarantees via Conformal Prediction

Debashish Chakraborty[1], Eugene Yang[1], Daniel Khashabi[2], Dawn Lawrie[1], and Kevin Duh[1]

[1] HLTCOE, Johns Hopkins University, Baltimore, Maryland, USA
{dchakra6,eugene.yang,lawrie,kduh1}@jhu.edu
[2] Johns Hopkins University, Baltimore, Maryland, USA
danielk@cs.jhu.edu
Correspondence: dchakra6@jhu.edu

Abstract. Retrieval-Augmented Generation (RAG) enhances factual grounding in large language models (LLMs) by incorporating retrieved evidence, but LLM accuracy declines when long or noisy contexts exceed the model's effective attention span. Existing pre-generation filters rely on heuristics or uncalibrated LLM confidence scores, offering no statistical control over retained evidence. We evaluate and demonstrate *context engineering through conformal prediction*, a coverage-controlled filtering framework that removes irrelevant content while preserving recall of supporting evidence. Using both embedding- and LLM-based scoring functions, we test this approach on the NeuCLIR and RAGTIME collections. Conformal filtering consistently meets its target coverage, ensuring that a specified fraction of relevant snippets are retained, and reduces retained context by 2–3× relative to unfiltered retrieval. On NeuCLIR, downstream factual accuracy measured by ARGUE F1 improves under strict filtering and remains stable at moderate coverage, indicating that most discarded material is redundant or irrelevant. These results demonstrate that conformal prediction enables reliable, coverage-controlled context reduction in RAG, offering a model-agnostic and principled approach to context engineering.

Keywords: Context Engineering · Retrieval Augmented Generation · Conformal Prediction.

1 Introduction

Retrieval-Augmented Generation (RAG) grounds large language models (LLMs) in retrieved evidence, reducing hallucinations compared to standalone models [15,26]. Despite rapid progress, RAG systems remain brittle, with retrieval noise and prompt saturation degrading reliability [21,3]. The *lost-in-the-middle* effect [18] shows that LLMs attend poorly to mid-prompt evidence, limiting effective use of long-context capacity to 10–20% [9,8]. Context has therefore been reframed as a finite *attention budget*, motivating high-signal, compact inputs for reliable generation [24]. Retrieval noise compounds this issue. Most vector

databases rank text by cosine similarity between dense embeddings, yet such similarity scores are typically uncalibrated and may be weakly correlated with true relevance [29]. Irrelevant or marginally related passages frequently enter the prompt, diluting useful evidence and inflating token costs. Benchmarks such as RAGTruth [23] and CRAG [35] show that such errors harm factual accuracy.

We address these challenges by introducing **context engineering through conformal prediction** as a principled mechanism for pre-generation filtering in RAG. Conformal prediction provides finite-sample coverage guarantees, ensuring that a specified proportion of relevant snippets are retained while irrelevant material is filtered out without additional model training [31,2]. Unlike prior RAG calibration methods that operate post-generation, our approach applies conformal prediction immediately after retrieval, offering formal control over context composition and context size.

We evaluate the framework on NeuCLIR [13] and RAGTIME [12]. Across both, conformal filtering achieves target coverage while reducing context size by 2–3×. On NeuCLIR, answer quality measured by ARGUE F1 [22] improves under strict filtering and remains stable at moderate levels, indicating that most removed content contributes little to downstream generation. Together, these results show that conformal prediction enables reliable, coverage-controlled context reduction, establishing it as a lightweight, model-agnostic foundation for principled context engineering in RAG. Our work makes three contributions:

1. We introduce a framework for **context engineering** in RAG, applying conformal prediction after retrieval to guarantee coverage of relevant evidence.
2. We empirically demonstrate that **conformal filtering** achieves target coverage while reducing context size by 2–3× across NeuCLIR and RAGTIME, maintaining factual accuracy under strict filtering.
3. We show that this approach is **model-agnostic**, needs no retraining, and works with both embedding- and LLM-based scoring functions.

2 Related Work

Prior research on mitigating retrieval noise in RAG systems can be grouped into three categories: heuristic filtering, LLM-based re-ranking, and conformal calibration. In this section, we review limitations of each class in turn.

Heuristic Filtering. Most production RAG pipelines rely on simple heuristics such as top-k retrieval or fixed similarity thresholds. Frameworks like LlamaIndex and vector databases such as Weaviate rank chunks by cosine distance [19,1]. While efficient, these methods may exhibit different effectiveness across topics [29]. Empirical studies find that irrelevant or marginally related snippets frequently pass such filters, diluting evidence and amplifying long-context degradation [23,11].
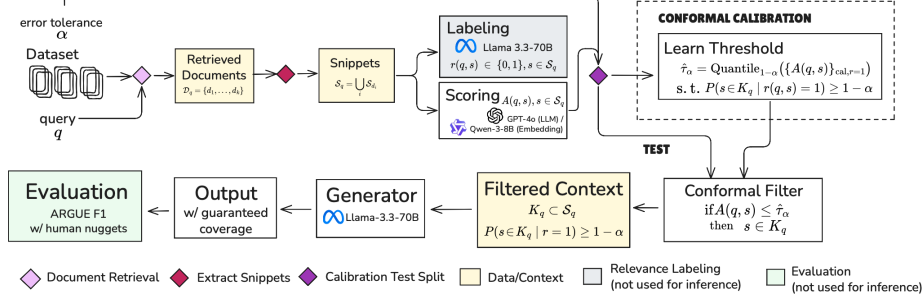
Fig. 1. Conformal context filtering workflow. Query q retrieves documents $\mathcal{D}_q$, segmented into snippets $\mathcal{S}_q$. Each snippet is scored $A(q,s)$, calibrated to threshold $\hat{\tau}_\alpha$, filtered to $K_q = \{s : A(q,s) \leq \hat{\tau}_\alpha\}$, and passed to generation. ARGUE F1 evaluates generated answers against human nuggets.

LLM-Based Filtering. Recent work has explored using the generator itself to assess snippet quality. LLatrieval prompts LLMs to judge retrieval sufficiency [17], MiniCheck employs LLMs for claim-level verification [30], and other pipelines decouple evidence selection from generation [27,28]. LLM confidence values are not probabilistic posteriors and are frequently miscalibrated – though monotonically correlated with relevance – they form coarse, prompt-sensitive scales that lack statistical calibration [7,34,33,20].

Conformal Prediction for RAG. Conformal prediction (CP) provides finite-sample, distribution-free guarantees on coverage without assuming a calibrated posterior [31,2]. Recent studies extend CP to RAG systems: **Conformal-RAG** improves group-conditional coverage for claim verification [6], while **C-RAG** bounds the risk of factual error during generation relative to standalone LLMs [11]. Closest to our setting, CONFLARE [25] calibrates similarity cutoffs to control retrieval uncertainty at the retrieval stage, while TRAQ [16] provides an end-to-end correctness guarantee over answer sets in retrieval-augmented Question Answering. Our contribution differs in scope: we apply split conformal prediction directly to snippet retention immediately after retrieval and evaluate its coverage efficiency trade-off (and downstream nugget-based factuality where available) under topic-disjoint calibration/test splits. By applying CP immediately after retrieval, we prevent noisy or redundant content from entering the generator, enabling on-the-fly filtering that constrains context length while ensuring coverage.

3 Methodology

3.1 Problem Formulation

Given a query q, a retriever returns a set of documents $\mathcal{D}_q = \{d_1, \ldots, d_k\}$ that may contain both relevant and irrelevant content, motivating snippet-level fil-

tering. Each document is segmented into 500-character windows overlapping by 100 characters total (50 on each side), preserving sentence boundaries following empirical chunking analysis [4]. Let $\mathcal{S}_q$ denote all retrieved snippets and $r(q,s) \in \{0,1\}$ indicate whether snippet s supports answering q. We aim to construct a filtered subset $K_q \subseteq \mathcal{S}_q$ that retains relevant snippets under a user-specified *miscoverage rate* $\alpha \in (0,1)$. Formally, the selection rule must achieve marginal coverage $P(s \in K_q \mid r(q,s) = 1) \geq 1 - \alpha$, ensuring that a labeled-relevant snippet (q,s) drawn exchangeably with calibration is retained with probability at least $1-\alpha$. A lower α provides stronger coverage guarantees at the cost of including more context, while higher α allows more aggressive filtering. Figure 1 summarizes the workflow.

3.2   Split Conformal Prediction for Context Filtering

We apply split conformal prediction [2,31,14] to obtain finite-sample marginal coverage guarantees. The method requires a **scoring function** $A(q,s)$ assigning nonconformity scores (lower = more relevant), a labeled **calibration set** $\mathcal{D}_{\mathrm{cal}}$, and a disjoint **test set** $\mathcal{D}_{\mathrm{test}}$ from the same distribution.

Given miscoverage α, the empirical $(1-\alpha)$-quantile of the positive calibration scores defines the filtering threshold:

$$\hat{\tau}_\alpha = \mathrm{Quantile}_{1-\alpha}\big(\{A(q,s) : (q,s) \in \mathcal{D}_{\mathrm{cal}}, r(q,s) = 1\}\big),$$

or equivalently, the $\lceil (n+1)(1-\alpha) \rceil$-th smallest score among n positive examples. At test time, a snippet s is retained iff $A(q,s) \leq \hat{\tau}_\alpha$. Under the exchangeability assumption between calibration and test splits, this guarantees $P(s \in K_q \mid r(q,s) = 1) \geq 1-\alpha$.

3.3   Experimental Setup

We now describe the experimental setup used to evaluate this framework. We evaluate on **NeuCLIR** [13] and **RAGTIME** [12], using disjoint query topics for calibration and test to preserve exchangeability (NeuCLIR: 1,440/740 snippets; RAGTIME: 1,710/560).

Scoring functions. We test two paradigms:

1. **Conformal-Embedding** using Qwen3-Embedding-8B [36], $A_{\mathrm{emb}}(q,s) = 1 - \cos(\mathrm{emb}(q), \mathrm{emb}(s))$; and
2. **Conformal-LLM** using GPT-4o [10] prompted to rate snippet relevance on $[0,1]$, $A_{\mathrm{LLM}}(q,s) = 1 - \mathrm{rating}$

Relevance labeling. Calibration and test relevance labels $r(q,s)$ are generated by Llama 3.3-70B-Instruct [5] using a rubric-style prompt, similar to [30], asking whether each snippet supports the query. The model outputs a binary decision parsed into $r(q,s) \in \{0,1\}$. Calibration labels define $\hat{\tau}_\alpha$; test labels are used only for empirical coverage evaluation. A 10% subsample was manually reviewed to verify consistency between human judgments and model labels.

Labeling as an annotation function. We treat the labeler as an *annotation function* that provides consistent binary relevance labels for conformal calibration and empirical coverage measurement. Human nuggets in NeuCLIR are used only for downstream evaluation (ARGUE F1) and are never used to set conformal thresholds. As with any automatically generated annotation, guarantees are conditional on label consistency across calibration and test topics. The exact prompt templates and output format used for LLM relevance scoring and labeling are released in our repository.[3]

Generation and evaluation. Filtered snippets K_q are concatenated and provided to the same Llama 3.3-70B generator for answer production to maintain consistency between labeling and generation. We report: (1) empirical coverage, (2) removal rate, and (3) downstream factual quality on NeuCLIR via ARGUE F1 [22] with the AutoARGUE implementation [32]. Since the nugget annotation of the RAGTIME collection, which is used for TREC RAGTIME Track in 2025, was not available at the time of conducting our experiments, it is therefore evaluated only for coverage and removal behavior.

4   Results and Discussion

Coverage Guarantees. Figure 2(a–b) shows empirical coverage against the target $(1 - \alpha)$ for **NeuCLIR** and **RAGTIME**. Across all α values (0.05–0.40), both Conformal-Embedding and Conformal-LLM meet or slightly exceed theoretical coverage guarantees, confirming the finite-sample validity of split conformal prediction. Conformal-LLM exhibits mild over-coverage (flat segments near 0.85–0.87) due to discretized rating bins, yielding coarser control over coverage levels. By contrast, Conformal-Embedding tracks the target line more smoothly, providing finer adaptation to coverage.

Heuristic pruning baselines. As a quantitative reference, per-query top-k snippet pruning yields uncontrolled coverage: on NeuCLIR, k=30 achieves 30% context reduction but only 76% empirical coverage, and k=25 achieves 39% reduction but only 68% coverage (w.r.t. $r(q, s)$), well below the 90–95% coverage levels targeted by conformal filtering.

Heuristic threshold baselines. Fixed cosine thresholds provide no coverage control and vary widely across queries: on NeuCLIR, θ=0.50 yields 76%±20% (mean ± std across queries) coverage at 35% context reduction, while θ=0.60 drops coverage to 43%±15% at 69% reduction. In contrast, conformal filtering targets a user-specified coverage level (e.g., 90%) and tracks it closely under exchangeability.
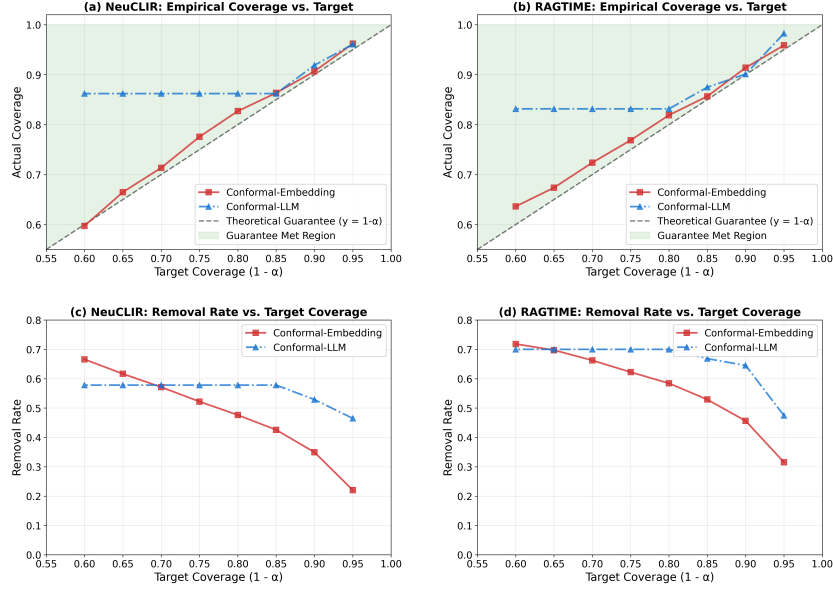
---

[3] https://github.com/hltcoe/conformal-context-engineering

Fig. 2. Coverage guarantees and context reduction across NeuCLIR and RAGTIME. (a–b) Empirical coverage vs. target $(1 - \alpha)$ (dashed: theoretical guarantee, shaded: valid region). Both methods satisfy conformal guarantees; Conformal-Embedding follows the target line more closely. (c–d) Removal rate vs. $(1-\alpha)$, illustrating the expected monotonic trade-off between tighter coverage and stronger filtering. Conformal-LLM removes more context overall but in quantized steps.

Context Reduction Efficiency. Figures 2(c–d) present removal rate as a function of target coverage. Removal decreases monotonically as $(1 - \alpha)$ increases, illustrating the expected trade-off between tighter guarantees and stronger filtering. At strict coverage ($\alpha \leq 0.20$), Conformal-Embedding removes 25–55% of retrieved snippets while maintaining full coverage, offering stable and interpretable control of context size. Conformal-LLM removes 46–70% of content but exhibits discrete jumps in removal rate due to its quantized confidence ratings. This consistent monotonic behavior across both datasets demonstrates that conformal filtering provides a reliable and tunable mechanism for managing retrieval depth.

Downstream Answer Quality. We assess the impact of filtering on factual generation quality on NeuCLIR using ARGUE F1 [22] with AutoARGUE [32]. The unfiltered baseline achieves 0.69 F1. Both filters improve ARGUE F1 at strict coverage (0.05–0.10) and remain indistinguishable from the baseline at α=0.20. Together with the coverage results, these findings show that conformal filtering effectively denoises retrieved context, removing redundant or weakly relevant snippets without harming factual accuracy. **Table 1** presents ARGUE F1 and

Table 1. NeuCLIR factual quality (ARGUE F1) and context reduction (ConRed%). Bold marks the best value per column. † indicates significant improvement over the unfiltered baseline ($p<0.05$, paired bootstrap resampling). The unfiltered baseline achieves 0.69 F1 with 0% reduction.

Method	α=0.05 (F1 / ConRed%)	α=0.10 (F1 / ConRed%)	α=0.20 (F1 / ConRed%)
Conformal-Embedding	**0.720**† / 22.2	**0.700**† / 35.0	0.680 / 52.8
Conformal-LLM	0.710† / 46.5	**0.700**† / 58.0	0.680 / 57.8

context reduction side by side for each α, highlighting the balance between factual retention and filtering efficiency.

Discussion. The results highlight three observations: **(1)** Split conformal prediction reliably enforces coverage guarantees across datasets and scoring models, turning pre-generation filtering from a heuristic into a statistically grounded process. **(2)** Conformal-Embedding provides smooth, predictable control at high coverage targets (80–95% marginal coverage of labeled-relevant snippets), whereas Conformal-LLM achieves stronger pruning but in coarse steps due to score quantization. **(3)** On NeuCLIR, ARGUE F1 improves at strict coverage ($\alpha \in \{0.05, 0.10\}$) and remains statistically indistinguishable from the baseline at α=0.20, indicating that more than half of retrieved snippets can be pruned without loss in factual quality. Although absolute gains are modest, the stability itself is informative: once retrieval noise is reduced, the generator likely operates near its effective attention limit. This finding reframes conformal prediction as a practical tool for *context engineering*, enabling robust, coverage-aware filtering before generation, laying the foundation for adaptive recalibration under domain or topic shifts.

5   Conclusion

We presented a statistical framework for *context engineering* in RAG based on split conformal prediction. Across NeuCLIR and RAGTIME, both embedding- and LLM-based conformal filters achieved guaranteed coverage while reducing context size by up to threefold. On NeuCLIR, downstream factual quality (ARGUE F1) improved under strict filtering and remained stable at moderate coverage, showing that redundant content can be safely pruned without loss of accuracy. These findings demonstrate that conformal prediction enables reliable, coverage-controlled context reduction and provides a lightweight, model-agnostic foundation for scalable RAG. Future work will explore adaptive recalibration across topics and domains to relax the exchangeability assumption and extend statistical guarantees under distribution shift.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Vector Indexing | Weaviate Documentation — docs.weaviate.io. https://docs. weaviate.io/weaviate/concepts/vector-index, [Accessed 24-09-2025]
2. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. ArXiv **abs/2107.07511** (2021), https: //api.semanticscholar.org/CorpusID:235899036
3. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Pimenta, D.: A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. medRxiv (2024). https://doi.org/10.1101/2024.09.12.24313556, https://www.medrxiv.org/content/early/2024/09/13/2024.09.12.24313556
4. Brådland, H., Olsen, M.G., Andersen, P.A., Nossum, A.S., Gupta, A.: A new hope: Domain-agnostic automatic evaluation of text chunking. ArXiv **abs/2505.02171** (2025), https://api.semanticscholar.org/CorpusID:278327433
5. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A.S., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., tian Cantón Ferrer, C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E.A., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I.M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J.Q., Alwala, K.V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K.R., El-Arini, K., Iyer, K., Malik, K., ley Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., hesh Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M.H.M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., lay Bashlykov, N., Bogoychev, N., Chatterji, N.S., Duchenne, O., cCelebi, O., Alrassy, P., Zhang, P., Li, P., Vasić, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., nie Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., hana Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S.C., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., ginie Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., ney Meers, W., Martinet, X., Wang, X., Tan, X.E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A.K., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A.,

Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, P.Y.B., Loyd, B., de Paola, B., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, S.W., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzm'an, F., Kanayet, F.J., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G.G., Zhang, G., Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Molybog, I., Tufanov, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., KamHou, U., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., Lavender, A., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., ish Bansal, M., Santhanam, N., Parks, N., White, N., ata Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N.P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., dro Rittner, P., Bontrager, P., Roux, P., Dollár, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Shankar, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S.K., Cho, S.B., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V.A., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Wang, Y., Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z.: The llama 3 herd of models. ArXiv **abs/2407.21783** (2024), https://api.semanticscholar.org/CorpusID:271571434

6. Feng, N., Sui, Y., Hou, S., Cresswell, J.C., Wu, G.: Response quality assessment for retrieval-augmented generation via conditional conformal factuality. ArXiv **abs/2506.20978** (2025), https://api.semanticscholar.org/CorpusID:280011519

7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. ArXiv **abs/1706.04599** (2017), https://api.semanticscholar.org/

CorpusID:28671436

8. Hong, K., Troynikov, A., Huber, J.: Context rot: How increasing input tokens impacts llm performance. Tech. rep., Chroma (July 2025), https://research.trychroma.com/context-rot

9. Hsieh, C.P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Ginsburg, B.: Ruler: What's the real context size of your long-context language models? ArXiv **abs/2404.06654** (2024), https://api.semanticscholar.org/CorpusID:269032933

10. Hurst, O.A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mkadry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., drey Mishchenko, A., Baek, A., Jiang, A., toine Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C.L., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C.J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mély, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, P.D., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F.P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H.B., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., woo Jun, H., Kirchner, H., de Oliveira Pinto, H.P., Ren, H., Chang, H., Chung, H.W., Kivlichan, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J.W., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J.R., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J.G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J.Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J.W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Long, O., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., laine Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., teusz Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M.A., Yang, M., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., Castro, M., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., talie Cone, N., Staudacher, N., Summers, N., LaFontaine, N.,

Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N.A., Felix, N., Kudige, N., Keskar, N.S., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., ing Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R.G., Puri, R., Miyara, R., Leike, R.H., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R.R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Xia, S., Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., Malkov, Y.: Gpt-4o system card. ArXiv **abs/2410.21276** (2024), https://api.semanticscholar.org/CorpusID:273662196

11. Kang, M., Gurel, N.M., Yu, N., Song, D.X., Li, B.: C-rag: Certified generation risks for retrieval-augmented language models. ArXiv **abs/2402.03181** (2024), https://api.semanticscholar.org/CorpusID:267412330

12. Lawrie, D., MacAvaney, S., Mayfield, J., Soldaini, L., Yang, E., Yates, A.: TREC RAGTIME: RAG TREC Instrument for Multilingual Evaluation. https://trec-ragtime.github.io (2025), official website for the TREC RAGTIME track

13. Lawrie, D.J., MacAvaney, S., Mayfield, J., McNamee, P., Oard, D.W., Soldaini, L., Yang, E.: Overview of the trec 2024 neuclir track (2025), https://api.semanticscholar.org/CorpusID:281394231

14. Lei, J., G'Sell, M.G., Rinaldo, A., Tibshirani, R.J., Wasserman, L.A.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**, 1094 – 1111 (2016), https://api.semanticscholar.org/CorpusID:13741419

15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021), https://arxiv.org/abs/2005.11401

16. Li, S., Park, S., Lee, I., Bastani, O.: TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3799–3821. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.210, https://aclanthology.org/2024.naacl-long.210/

17. Li, X., Zhu, C., Li, L., Yin, Z., Sun, T., Qiu, X.: LLatrieval: LLM-verified retrieval for verifiable generation. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 5453–5471. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.305, https://aclanthology.org/2024.naacl-long.305/

18. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024). https://doi.org/10.1162/tacl_a_00638, https://aclanthology.org/2024.tacl-1.9/

19. LlamaIndex Developers: Llamaindex python framework: Embeddings module guide. https://developers.llamaindex.ai/python/framework/module_guides/models/embeddings/ (2025), accessed: October 2025

20. Lovering, C., Krumdick, M., Lai, V.D., Ebner, S., Kumar, N., Reddy, V., Koncel-Kedziorski, R., Tanner, C.: Language model probabilities are not calibrated in numeric contexts (2024), https://api.semanticscholar.org/CorpusID:273502432

21. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C.D., Ho, D.E.: Hallucination-free? assessing the reliability of leading ai legal research tools. ArXiv **abs/2405.20362** (2024), https://api.semanticscholar.org/CorpusID:269976547

22. Mayfield, J., Yang, E., Lawrie, D., MacAvaney, S., McNamee, P., Oard, D.W., Soldaini, L., Soboroff, I., Weller, O., Kayi, E., Sanders, K., Mason, M., Hibbler, N.: On the evaluation of machine-generated reports. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1904–1915. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3657846, https://doi.org/10.1145/3626772.3657846

23. Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., Zhang, T.: RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10862–10878. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.585, https://aclanthology.org/2024.acl-long.585/

24. Rajasekaran, P., Dixon, E., Ryan, C., Hadfield, J., Ayub, R., Moran, H., Rueb, C., Jennings, C.: Effective context engineering for ai agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents (2025), anthropic Engineering Blog, Published September 29, 2025

25. Rouzrokh, P., Faghani, S., Gamble, C., Shariatnia, M., Erickson, B.J.: Conflare: Conformal large language model retrieval. ArXiv **abs/2404.04287** (2024), https://api.semanticscholar.org/CorpusID:269004787

26. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval augmentation reduces hallucination in conversation. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 3784–3803. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.320, https://aclanthology.org/2021.findings-emnlp.320/

27. Slobodkin, A., Hirsch, E., Cattan, A., Schuster, T., Dagan, I.: Attribute first, then generate: Locally-attributable grounded text generation. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3309–3344. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.182, https://aclanthology.org/2024.acl-long.182/

28. Sohn, J., Park, Y., Yoon, C., Park, S., Hwang, H., Sung, M., Kim, H., Kang, J.: Rationale-guided retrieval augmented generation for medical question answering. ArXiv **abs/2411.00300** (2024), https://api.semanticscholar.org/CorpusID:273798271

29. Steck, H., Ekanadham, C., Kallus, N.: Is cosine-similarity of embeddings really about similarity? Companion Proceedings of the ACM Web Conference 2024 (2024), https://api.semanticscholar.org/CorpusID:268296965

30. Tang, L., Laban, P., Durrett, G.: MiniCheck: Efficient fact-checking of LLMs on grounding documents. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 8818–8847. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.499, https://aclanthology.org/2024.emnlp-main.499/

31. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer-Verlag, Berlin, Heidelberg (2005)

32. Walden, W.G., Mason, M., Weller, O., Dietz, L., Recknor, H., Li, B., Liu, G.K.M., Hou, Y., Mayfield, J., Yang, E.: Auto-argue: Llm-based report generation evaluation (2025), https://api.semanticscholar.org/CorpusID:281682210

33. Xie, Q., Li, Q., Yu, Z., Zhang, Y., Zhang, Y., Yang, L.: An empirical analysis of uncertainty in large language model evaluations. ArXiv **abs/2502.10709** (2025), https://api.semanticscholar.org/CorpusID:276408437

34. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., Hooi, B.: Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. ArXiv **abs/2306.13063** (2023), https://api.semanticscholar.org/CorpusID:259224389

35. Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R.D., Jiang, Z.W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y.E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., Chen, L., Scheffer, N., Liu, Y., Shah, N., Wanga, R., Kumar, A., Yih, W.t., Dong, X.L.: Crag - comprehensive rag benchmark. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. NIPS '24, Curran Associates Inc., Red Hook, NY, USA (2025)

36. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. ArXiv **abs/2506.05176** (2025), https://api.semanticscholar.org/CorpusID:279243736