KR Labs at ArchEHR-QA 2025: A Verbatim Approach for Evidence-Based Question Answering

Ádám Kovács[1], **Paul Schmitt**[2], **Gábor Recski**[1,2]

[1]KR Labs
`lastname@krlabs.eu`

[2]TU Wien
`firstname.lastname@tuwien.ac.at`

Abstract

We present a lightweight, domain-agnostic *verbatim* pipeline for evidence-grounded question answering. Our pipeline operates in two steps: first, a sentence-level extractor flags relevant note sentences using either zero-shot LLM prompts or supervised ModernBERT classifiers. Next, an LLM drafts a question-specific template, which is filled verbatim with sentences from the extraction step. This prevents hallucinations and ensures traceability. In the ArchEHR-QA 2025 shared task, our system scored 42.01%, ranking top-10 in core metrics and outperforming the organiser's 70B-parameter Llama-3.3 baseline. We publicly release our code and inference scripts under an MIT license.

1 Introduction

Modern question-answering (QA) and retrieval-augmented generation (RAG) systems play a vital role in many high-stakes domains for information extraction and generation tasks. In medicine, a typical use case involves clinicians asking questions based on a patient's electronic health record (EHR) notes, rather than manually sifting through lengthy notes, which can be time-consuming. However, in practice, RAG and QA pipelines often misalign evidence and produce incorrect information, commonly referred to as hallucinations (Ji et al., 2023; Madsen et al., 2024). We argue that a reliable QA system should guarantee complete traceability of answers. To tackle this problem, we propose a *verbatim* pipeline that clearly separates extraction and generation to mitigate hallucinations:

- **Sentence-level extraction**, using either zero-shot LLMs or supervised ModernBERT classifiers.
- **Template-constrained generation**, dynamically creating answer templates filled exclusively with verbatim sentences selected from the extraction phase.

We participated in the ArchEHR-QA 2025 shared task on grounded question answering (QA) from electronic health records (EHRs). Our approach involved (i) utilizing a zero-shot `gemma-3-27b-it`[1] LLM (Team et al., 2025) and (ii) generating synthetic data for sentence extraction from EHRs to train a compact extractor. For this purpose, we trained a Clinical ModernBERT classifier (Lee et al., 2025; Warner et al., 2024), achieving performance comparable to the LLM extractor. Both extractors were then fed into the same LLM template generator. Our solution achieved an overall score of **42.01%**, ranking in the **top 10** for core metrics, and surpassed the organizers' 70B-parameter Llama-3.3 baseline by a large margin.

Our contributions include a modular, traceable QA architecture that mitigates hallucinations, a method to generate synthetic EHR question-answer corpus and train custom models. Additionally, we are releasing all the code on GitHub[2] under the MIT License. The remainder of the paper discusses background (Section 2), method (Section 3), and evaluation (Section 4).

2 Background

2.1 Dataset

Early clinical QA datasets such as emrQA (Pampari et al., 2018) and CliCR (Šuster and Daelemans, 2018) used fill-in-the-blank methods and lacked explicit sentence-level evidence. ArchEHR-QA (Soni and Demner-Fushman, 2025b,a) addresses this by pairing clinician-authored questions with de-identified MIMIC-III (Johnson et al., 2016) notes, annotated at the sentence-level as *essential*, *supplementary*, or *irrelevant*. Answers must be concise

[1] `https://huggingface.co/google/gemma-3-27b-it`

[2] `https://github.com/KRLabsOrg/verbatim-rag/tree/archehr`

BioNLP 2025 Shared Tasks, pages 69–74
August 1, 2025 ©2025 Association for Computational Linguistics

(under 75 words) and explicitly cite relevant sentences.

2.2 Limitations of Standard RAG

Standard RAG models, despite external grounding, still frequently hallucinate unsupported or contradictory information (Ji et al., 2023). Existing approaches like post-hoc verification (Friel and Sanyal, 2023; Manakul et al., 2023) or classifiers trained on hallucination corpora such as RAGTruth (Niu et al., 2024) (e.g., RAG-HAT (Song et al., 2024), LettuceDetect (Ádám Kovács and Recski, 2025)) add extra complexity and latency. Post-hoc saliency methods (Serrano and Smith, 2019; Jain and Wallace, 2019) and LLM self-explanations (Madsen et al., 2024) have also been found unreliable. Our approach proactively prevents hallucinations through strict template-driven sentence extraction and verbatim insertion.

2.3 Synthetic Training Data

Due to limited access and annotation restrictions, obtaining sentence-level labeled clinical datasets is challenging. Recent works address this by generating synthetic data via perturbation or LLM prompting (Niu et al., 2024; Lozano et al., 2023; Frayling et al., 2024; Bai et al., 2024). We follow this approach, generating synthetic EHR snippets, clinician-style questions, and sentence relevance annotations (details in Section 3.3).

3 Method

3.1 System Overview

Figure 1 depicts our system architecture. First, an extraction step identifies relevant sentences from the input (patient narrative, clinician question, and note excerpt). We implemented both zero-shot and supervised models. Second, the generation step uses `gemma-3-27b-it` to dynamically draft an answer template, filled verbatim with extracted sentences. If exceeding 75 words, answers are compressed via a summarization prompt, preserving sentence-level citations.

3.2 Evidence Extraction

We evaluated two extractors: (i) We prompted `gemma-3-27b-it` to explicitly label sentences as relevant via a step-by-step process. (ii) We fine-tuned a Clinical ModernBERT classifier (Lee et al., 2025), trained on our synthetic data (Section 3.3). It independently evaluates each sentence in context (question + patient narrative). Lee et al. (2025) is a variant of ModernBERT (Warner et al., 2024) adapted specifically for biomedical and clinical text. Clinical ModernBERT supports extended input sequences (up to 8,192 tokens) and includes domain-specific vocabulary enhancements, making it particularly suitable for handling long clinical narratives. To provide additional context during classification, we included one sentence before and after the target sentence, forming a passage of up to three sentences. We chose a window size of one sentence before and after the target based on preliminary experimentation. The target sentence was explicitly marked with `[START]` and `[END]` tokens. The full input was structured using the standard BERT classification format. During fine-tuning, we merged essential and supplementary labels into a single positive class. We addressed class imbalance using weighted binary cross-entropy loss. We trained for 3 epochs (batch size 32, learning rate 2e-5), with gradient clipping and early stopping based on F1 score.

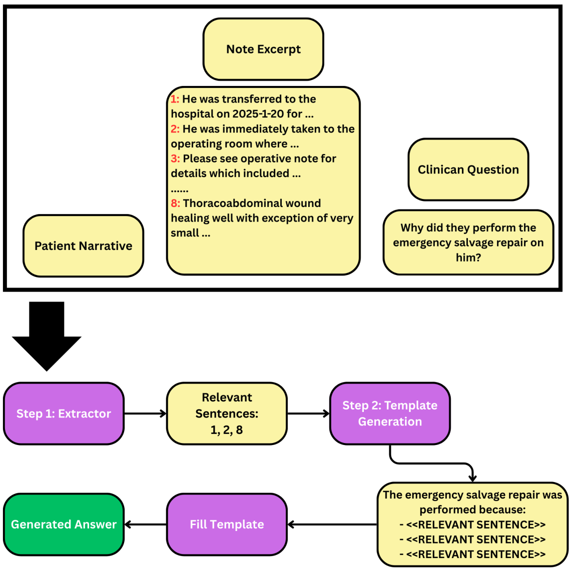
Note Excerpt
1: He was transferred to the hospital on 2025-1-20 for ...
2: He was immediately taken to the operating room where ...
3: Please see operative note for details which included ...
......
8: Thoracoabdominal wound healing well with exception of very small ...
Patient Narrative
Clinican Question
Why did they perform the emergency salvage repair on him?
Step 1: Extractor
Relevant Sentences: 1, 2, 8
Step 2: Template Generation
The emergency salvage repair was performed because:
- <<RELEVANT SENTENCE>>
- <<RELEVANT SENTENCE>>
- <<RELEVANT SENTENCE>>
Fill Template
Generated Answer

Figure 1: System overview. The pipeline first selects relevant sentences and then generates a question-specific answer using a dynamic template.

3.3 Synthetic Data Generation

Due to the scarcity of publicly available annotated data for sentence-level relevance classification, we constructed a synthetic dataset tailored specifically to the ArchEHR-QA task. Although the official development set contains labeled sentences, it is limited to 428 sentences across only 20 question–note pairs. Initial experiments using external resources like RAGBench (Friel et al., 2025) and PubMedQA-derived corpora (Jin et al., 2019) showed poor transfer performance, emphasizing the need for task-specific synthetic data.

We generated synthetic data via few-shot prompting with `gemma-3-27b-it`. Each prompt provided dynamic examples from the development set to ensure diversity. The LLM generated synthetic instances comprising de-identified clinical note excerpts, patient narratives, clinician-authored questions, and binary relevance labels. This approach yielded **3915** synthetic notes. We varied the few-shot examples across multiple runs, as static prompting resulted in repetitive outputs. This variation greatly increased lexical and semantic diversity, aligning with other work in synthetic data generation (Li et al., 2023; Tang et al., 2023; Xu et al., 2024). Ultimately, selecting each sentence with their relevance from the note excerpts, we constructed a comprehensive dataset of **58k** synthetic training examples, each labeled at the sentence level, which formed the training set for our Clinical ModernBERT classifier. Table 1 shows an illustrative training instance.

QUESTION	**Patient narrative**: My husband, a 72-year-old with a history of COPD, was admitted for worsening shortness of breath. He's been on home oxygen for years, but it wasn't helping this time. He also developed some swelling in his ankles. He seems a little confused today... **Clinician question**: What is the likely cause of the patient's ankle edema and what was done to address it?
SENTENCE	A diuretic, furosemide 40mg PO daily, was initiated to address the lower extremity edema, which was attributed to both underlying heart failure and fluid retention secondary to COPD exacerbation. **[START]** Echocardiogram revealed mild left ventricular dysfunction with an estimated ejection fraction of 45%. **[END]** Renal function was monitored closely, and remained stable throughout hospitalization.
LABEL	RELEVANT

Table 1: An example model input for our training.

3.4 Answer Generation

The answer generation module dynamically creates a template using the LLM (`gemma-3-27b-it`) based on the clinician's question, the selected evidence sentences, and the clinical note context. After the template generation step, we directly insert the extracted evidence sentences verbatim into the generated template, referencing sentence IDs explicitly. An example filled template generated by our pipeline is shown in Figure 2.

```
The emergency salvage repair was performed due to:
- He was transferred to the hospital on 2025-01-20 for emergent
repair of his ruptured thoracoabdominal aortic aneurysm. |1|
- He was immediately taken to the operating room where he underwent an
emergent salvage repair of ruptured thoracoabdominal aortic aneurysm with
a 34-mm Dacron tube graft using deep hypothermic circulatory arrest. |2|
- Thoracoabdominal wound healing well with exception of very small open
area mid-wound that is ~1 cm around and 0.5 cm deep, no surrounding
erythema. |8|
```

Figure 2: Example answer generated by our *verbatim* method, inserting evidence sentences verbatim into a dynamically generated template.

```
He was transferred to the hospital on 2025-01-20 for emergent repair of
      his ruptured thoracoabdominal aortic aneurysm |1|. He underwent an
       emergent salvage repair with a 34-mm Dacron tube graft using deep
       hypothermic circulatory arrest |2|. See also: |8|
```

Figure 3: Concise answer produced by our summarization step to comply with the 75-word limit.

If the filled answer exceeds the 75-word constraint of the task, we use an additional summarization prompt to rewrite the answer more concisely, ensuring all selected evidence remains cited and intact. An example summarization of the answer from Figure 2 is illustrated in Figure 3.

4 Evaluation

We evaluated our pipeline in the ArchEHR-QA 2025 shared task (Soni and Demner-Fushman, 2025b) using official metrics that emphasize two main aspects. Factuality is measuring alignment of the cited evidence with manually annotated sentences. Citation-level F1 scores are computed under strict (essential sentences only) and lenient (essential and supplementary sentences) conditions. Relevance is evaluating how closely generated answers match ground-truth answers through BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and MEDCON (Yim et al., 2023) metrics.

Our best system (zero-shot LLM based on `gemma-3-27b-it`) scored **42.01%** overall, placing within the top 10 in multiple core metrics and significantly outperforming the organizers' baseline (a 70B-parameter Llama-3.3 model) across most metrics. Table 2 summarizes these metrics. A strong point of our system is factuality recall (56.8% strict, 56.6% lenient), approximately 5 points above the leaderboard average. This indicates strong capability for reliably retrieving relevant clinical evidence. Precision was more moderate (48.1% strict, 50.7% lenient), suggesting that our methods are more recall oriented in the extraction phase. In terms of

Table 2: ArchEHR-QA 2025 test-set scores: our zero-shot LLM system (*KR-Labs* versus the organizer baseline, Llama-3.3-70B).

Team	**Ov.**	**Fact.**	**Rel.**	**Strict μ**			**Len. μ**			**Strict M**			**Len. M**			**BLEU**	**R-L**	**SARI**	**BERT**	**Align**	**MEDC**
				P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**						
KR-Labs	42.0	52.1	31.9	48.1	56.8	52.1	50.7	56.6	53.5	55.8	62.3	54.3	60.4	60.6	56.2	2.0	21.4	57.9	26.3	49.0	35.2
Organizer baseline	30.7	33.6	27.8	71.6	21.9	33.6	77.0	22.3	34.6	77.4	31.5	39.0	83.0	30.8	39.9	0.1	15.2	47.8	20.5	57.7	25.6

relevance, our system achieved competitive AlignScore (49.0%) and MEDCON (35.2%).

We compared our zero-shot LLM extractor and the fine-tuned Clinical ModernBERT extractor on the development set, these findings can be seen in Table 3. The comparison highlights a clear tradeoff: the LLM-based extractor provides higher precision and balanced F1, while the Clinical ModernBERT demonstrates strong recall, capturing nearly all relevant information at the expense of precision. Our final submission employed the LLM extractor for its balanced performance.

Table 3: Sentence-level extraction on the development set.

Extractor	**Precision**	**Recall**	**F1**
LLM (gemma-3-27b-it)	0.56	0.73	**0.63**
Clinical ModernBERT	0.46	**0.91**	0.61

Interestingly, final test scores were closely matched between our extractors: the LLM-based model scored 42.01%, while Clinical ModernBERT achieved a near-identical 41.85%. This underscores the effectiveness of our synthetic data training methodology, enabling a lightweight model to achieve comparable performance to a larger LLM.

Overall, our results demonstrate that lean methods can achieve competitive performance in EHR QA, highlighting the value of synthetic data generation. We show that even smaller LLMs, when used for data creation, can enable the training of lightweight models that rival larger systems—while requiring significantly fewer computational resources.

5 Ethical Considerations

Our experiments were conducted exclusively on a secure, private A100 GPU server. This ensured that we adhered to all data licensing requirements and maintained confidentiality throughout the project lifecycle, making the data inaccessible externally. Our work relies on de-identified clinical text and the generation of synthetic data. However, it is important to note that clinical AI systems can perpetuate harmful biases (Bender et al., 2021; Obermeyer et al., 2019). In any deployment setting, we recommend implementing a human-in-the-loop review process, maintaining strict provenance tracking of cited evidence, and conducting thorough bias audits to ensure patient safety and fairness.

6 Limitations

Our *verbatim* RAG pipeline explicitly cites source sentences to mitigate hallucinations; however, several practical limitations remain. Due to the task's strict 75-word limit, our approach often required summarization after the initial verbatim insertion step, meaning the purely verbatim property was not consistently maintained across all answers. Additionally, although extracted sentences were cited exactly, the dynamically generated templates themselves were produced by an LLM, potentially introducing subtle hallucinations or inaccuracies at the framing level. Future work should include explicit checks on template factuality. Finally, user studies and clinician feedback are essential to confirm whether our structured, template-based answers effectively address real-world clinician information needs.

7 Conclusion

In this paper we presented a lightweight and transparent *verbatim* pipeline for grounded question answering from clinical texts. Our method separates sentence-level extraction from template-based generation, significantly reducing hallucinations and maintaining traceable evidence. Participating in the ArchEHR-QA 2025 shared task, our system ranked among the top-10 submissions on key metrics and significantly outperformed a substantially larger baseline (70B-parameter Llama-3.3). We also demonstrated the effectiveness of synthetic training data generated by smaller LLMs for developing competitive, resource-efficient models.

References

Fan Bai, Keith Harrigian, Joel Stremmel, Hamid Hassanzadeh, Ardavan Saeedi, and Mark Dredze. 2024. Give me some hard questions: Synthetic data generation for clinical qa. *Preprint*, arXiv:2412.04573.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmuel Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Erlend Frayling, Jake Lever, and Graham McDonald. 2024. Zero-shot and few-shot generation strategies for artificial clinical records. *arXiv preprint arXiv:2403.08664*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint*, arXiv:2407.11005.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *Preprint*, arXiv:2310.18344.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint*, arXiv:1909.06146.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. 2025. Clinical modernbert: An efficient and long context encoder for biomedical text. *Preprint*, arXiv:2504.03964.

Rumeng Li, Xun Wang, and Hong Yu. 2023. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7129–7143.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo.ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. *Preprint*, arXiv:2310.16146.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

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, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

Sarvesh Soni and Dina Demner-Fushman. 2025a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *arXiv preprint*.

Sarvesh Soni and Dina Demner-Fushman. 2025b. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Simon Šuster and Walter Daelemans. 2018. CliCR: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563, New Orleans, Louisiana. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D. Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. *arXiv preprint arXiv:2404.18443*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, Meliha Yetisgen, and *et al.* 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Ádám Kovács and Gábor Recski. 2025. LettuceDetect: A hallucination detection framework for RAG applications. *Preprint*, arXiv:2502.17125.