Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 502-520
DOI: https://doi.org/10.51903/jtie.v4i2.536

# Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0

Ziliang Samuel Zhong[1], Jing Chen*[2], Eric Zhong[3], Xinzhuo Sun[4]
Email: jingc0606@gmail.com
[1]*New York University, NY, USA*
[2]*Industrial Engineering and Operations Research, UCB, CA, USA*
[3]*Computer Science, USC, CA, USA*
[4]*Computer Engineering, Cornell Tech, NY, USA*
*Corresponding Author

**Abstract**
*This paper presents a controlled and reproducible empirical study of evidence-calibrated retrieval-augmented question answering (RAG) for answerable and unanswerable reading-comprehension tasks using the SQuAD 2.0 benchmark. The study focuses on whether a system should abstain when retrieved evidence is insufficient rather than always producing an answer. Six lightweight architectures were evaluated on the full validation set of 11,873 questions, including closed-book, BM25, dense, hybrid, reranked, and a proposed evidence-calibrated hybrid RAG model. The proposed approach combines hybrid top-25 retrieval, lexical reranking, deterministic extractive answering, and evidence sufficiency calibration trained on 43,482 examples. On the validation set, it achieved 31.65% exact match, 34.74% F1, 53.01% answerability accuracy, 53.71% refusal F1, and a 37.49% hallucination-proxy rate. Although overall QA performance remains modest, calibrated evidence sufficiency substantially reduced unsupported answers compared with a forced-answer hybrid reranker, lowering the hallucination-proxy rate from 77.80% while improving F1. However, evidence calibration itself remained weak (AUROC 0.5475, ECE 0.1144). The findings demonstrate that retrieval coverage alone is insufficient to prevent hallucinations and highlight the need for stronger evidence calibration in trustworthy RAG systems.*

**Keywords**: *retrieval-augmented generation; unanswerable question answering; abstention calibration; evidence sufficiency; hallucination-proxy reduction*

## I. INTRODUCTION

Question answering systems are often evaluated by how often they return the correct span, but real use requires a second ability: the system must know when not to answer. This requirement is central for retrieval-augmented generation (RAG), where an answer can be fluent even when the retrieved evidence is irrelevant, incomplete, or contradictory. RAG is attractive because retrieval supplies non-parametric evidence that complements parametric language-model knowledge (Lewis et al., 2020; Karpukhin et al., 2020). However, retrieval itself does not guarantee that the answer is supported by the evidence. A system can retrieve the wrong paragraph, retrieve a partially relevant paragraph, or retrieve a paragraph that contains overlapping terms but does not support the asked proposition.

SQuAD 2.0 operationalizes this problem directly. The benchmark combines answerable SQuAD-style span questions with adversarial unanswerable questions that resemble answerable ones (Rajpurkar et al., 2016; Rajpurkar et al., 2018). In this setting, an evidence-grounded model must return a span only when the paragraph supports it and must abstain when no supported answer is

*Received on May 2025; Revised on Juni 2025; Accepted on July 2025; Published on August 2025*

present. This is stricter than ordinary extractive QA because the evaluation rewards both correct answering and correct refusal. It is also a good stress test for RAG (Sun et al., 2024) because a retriever can find a semantically related paragraph while the answerer still produces a false answer.

The present study examines whether a compact, fully reproducible RAG pipeline can convert retrieval evidence into calibrated abstention decisions. The work is a controlled empirical evaluation rather than a state-of-the-art SQuAD 2.0 system: every reported number in the results tables is computed from a full evaluation on the SQuAD 2.0 validation split. The training split is used for question-type priors and abstention calibration, while all six systems are evaluated on the same validation questions and paragraph corpus. The proposed system separates three functions that are often conflated in RAG papers: retrieval coverage, answer extraction, and evidence sufficiency. Retrieval coverage asks whether the gold paragraph is found; answer extraction asks whether a span-like answer is produced; evidence sufficiency asks whether the evidence is strong enough for the system to answer rather than refuse (Zhou et al., 2023).

The motivation for this design is that many RAG failures are not failures of language fluency. A generated answer can be grammatical, plausible, and topically related while still being unsupported by the retrieved paragraph. In ordinary open-domain settings this problem is difficult to audit because the correct evidence may exist elsewhere. SQuAD 2.0 is stricter and cleaner: the question is evaluated against a known paragraph, and a system that answers without support is explicitly wrong. This allows the experiment to measure a concrete form of faithfulness rather than relying only on subjective judgments of hallucination.

Three research questions guide the experiment. RQ1 asks how retrieval coverage differs across sparse, dense, hybrid, and reranked retrieval on the SQuAD 2.0 validation corpus. RQ2 asks whether high retrieval coverage produces high end-to-end QA performance when unanswerable questions are included. RQ3 asks whether a calibrated evidence-sufficiency threshold reduces unsupported answers while preserving useful answer behavior. These questions are answered with the same dataset split, evaluator, and deterministic code so that the tables and figures can be reproduced from the included package.

The contribution is therefore methodological and diagnostic, not a claim of competitive QA performance. Methodologically, the paper defines a reproducible evidence-calibrated RAG pipeline that uses the same retrieved paragraphs for a forced-answer system and for a calibrated abstaining system. Diagnostically, it reports a measurement suite that separates retrieval coverage, answerability classification, refusal quality, calibration error, and hallucination-proxy behavior. This separation matters because a model can improve one axis while degrading another. For

example, a retriever can increase top-k coverage while a reader still over-answers unanswerable questions, and a strict abstention threshold can reduce unsupported answers while creating false refusals. The experiment is designed to expose these trade-offs rather than compress them into a single score.

## II. LITERATURE REVIEW

Open-domain and reading-comprehension QA have long relied on retrieval followed by reader models. DrQA showed that sparse retrieval plus a neural reader can answer many Wikipedia questions (Chen et al., 2017). BM25 remains a strong and interpretable sparse retrieval baseline because it weights term frequency, document length, and inverse document frequency in a transparent way (Robertson & Zaragoza, 2009). Neural dense retrieval later improved semantic matching by learning vector representations for questions and passages, as shown by dense passage retrieval (Karpukhin et al., 2020), sentence-transformer representations (Reimers & Gurevych, 2019), and broader retrieval benchmarks such as BEIR (Thakur et al., 2021). Reranking is also important because the first-stage retriever may find a broad candidate set while a reranker selects the final evidence (Nogueira & Cho, 2019).

RAG connects retrieval with generation. Lewis et al. (2020) formulated retrieval-augmented generation as a sequence-to-sequence model conditioned on retrieved passages. Fusion-in-Decoder showed that aggregating multiple retrieved passages can improve knowledge-intensive generation (Izacard & Grave, 2021). Recent work also emphasized self-reflection or self-retrieval, where a system explicitly evaluates whether retrieval is needed or whether its generated output is supported (Asai et al., 2023). These developments are relevant to answer abstention because the system must not treat retrieval as a license to answer; it must evaluate whether the retrieved evidence actually supports the answer.

Hallucination and faithfulness research has shown that fluent generation can diverge from source evidence. Maynez et al. (2020) documented factual inconsistency in abstractive summarization, and Ji et al. (2023) surveyed hallucination in natural language generation. For QA, adversarial examples can expose failures that are hidden by superficial lexical overlap (Jia & Liang, 2017). Menick et al. (2022) argued that language models should support answers with evidence, and Mallen et al. (2023) examined when parametric memory and retrieval should be trusted. These studies motivate the present focus on evidence sufficiency rather than answer fluency.

Calibration and selective prediction provide the statistical foundation for abstention. Guo et al. (2017) showed that modern neural networks are often miscalibrated, while Ovadia et al. (2019) showed that calibration can deteriorate under distribution shift. Hendrycks and Gimpel (2017) studied confidence baselines for detecting incorrect predictions, and Jiang et al. (2021) examined

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 4 No. 2 August 2025**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 502-520
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

whether language models know when they know. Kamath et al. (2020) applied selective prediction to question answering under domain shift. In unanswerable QA, calibration is operational: a threshold must decide whether the system outputs a span or abstains.

Unanswerable QA extends the older TREC and open-domain QA tradition by making refusal part of the task (Voorhees, 1999). In SQuAD 2.0, many unanswerable questions are written to look answerable, which means surface lexical similarity can be misleading (Rajpurkar et al., 2018). This adversarial property is important for RAG (Kuo et al., 2025) because retrieval modules often reward lexical overlap. A question and paragraph can share named entities and keywords while the paragraph still lacks the requested relation, date, cause, or quantity. Evidence sufficiency scoring is therefore a natural complement to retrieval scoring.

Faithfulness research further motivates an error taxonomy. Hallucination is not a single phenomenon: an answer may be unsupported because retrieval missed the paragraph, because the answerer chose the wrong sentence, because the system over-generalized from a partial clue, or because it should have refused an unanswerable item (Ji et al., 2023; Maynez et al., 2020). Reporting only EM and F1 compresses these different errors into one score. The present paper therefore reports retrieval coverage, answerability confusion, calibration bins, hallucination-proxy rates, and categorized errors (Zheng et al., 2024).

Prior datasets also clarify why SQuAD 2.0 is a suitable benchmark for the present question. TriviaQA and Natural Questions evaluate large-scale reading comprehension under more open retrieval conditions, but their evidence structure makes refusal analysis less controlled (Joshi et al., 2017; Kwiatkowski et al., 2019). SQuAD 1.1 provided high-quality paragraph-grounded span annotations, yet it did not require systems to decide that a paragraph lacks an answer (Rajpurkar et al., 2016). SQuAD 2.0 added adversarial unanswerable items to this format, giving a clean test of whether a model can distinguish topical relevance from answer support (Rajpurkar et al., 2018). For RAG evaluation, this distinction is essential because retrieval often supplies topically relevant text before the answerer decides whether the text is sufficient.

## III. RESEARCH METHOD

Dataset. The experiment used SQuAD 2.0. The parsed data preserved one question-level record per item and one paragraph-level record per retrieval context. The final expected split sizes were 130,319 training questions and 11,873 validation questions. The validation split contained 5,928 answerable questions, 5,945 unanswerable questions, and 1,204 unique validation paragraphs. These counts match the known SQuAD 2.0 split sizes and support full-split evaluation. Table 1 summarizes the parsed dataset statistics used for training calibration and validation evaluation. The reproducibility package includes the raw Wolfram CSV dataset, source code for generating

processed validation files, generated tables, generated figures, and experiment summary metadata.

Table 1. Dataset statistics after parsing the SQuAD 2.0 release

| Split | Questions | Answerable | Unanswerable | Paragraphs | Titles | Mean context tokens |
|---|---|---|---|---|---|---|
| Training | 130319 | 86821 | 43498 | 19035 | 442 | 128.91 |
| Validation | 11873 | 5928 | 5945 | 1204 | 35 | 128.91 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

Experimental systems. Six systems were evaluated. Table 2 summarizes their retrieval, reader, evidence-sufficiency, and abstention components. The No-RAG baseline was operationalized as a closed-book question-type prior generator that used the training split to choose the most frequent answer for each question type and had no access to retrieved evidence. BM25-RAG used a BM25 index over the validation paragraphs and applied the same deterministic extractive answerer to the top paragraph. Dense Retrieval RAG used deterministic dense-hash embeddings, where each token was mapped to a stable signed vector and contexts were represented by inverse-document-frequency weighted averages. This dense component was not a trained dense retriever; it was included as a reproducible approximation for controlled comparison. BM25 + Dense Hybrid RAG combined min-max normalized BM25 and dense scores. Hybrid RAG + Reranker reranked the top 25 hybrid candidates using sentence-overlap, score-fusion, and answer-type cues. The proposed system used the same reranked evidence but added evidence sufficiency scoring and calibrated abstention. Accordingly, the evaluated systems should be interpreted as controlled reproducible baselines, not as representatives of modern production RAG systems using trained dual encoders, cross-encoder rerankers, or LLM readers. Figure 1 gives an overview of the proposed evidence-calibrated RAG pipeline.

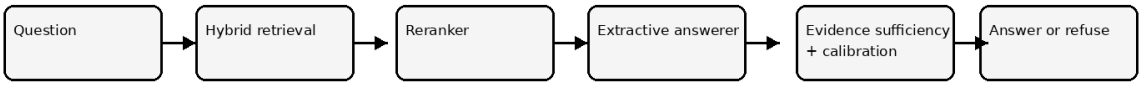


Figure 1. Evidence-calibrated RAG pipeline

Evidence sufficiency and abstention calibration. The sufficiency score combined question-context coverage, best-sentence coverage, best-sentence Jaccard overlap, question-type evidence cues, answer-extraction confidence, and title overlap. The threshold was calibrated on 43,482 deterministic training examples selected by a stable hash of the question ID. The safety-weighted calibration objective placed more weight on refusal quality than on answer recall because the research question concerns unsupported answering. The selected threshold was 0.59. At inference time, the proposed system returned the extracted answer only when the retrieved paragraph reached this threshold; otherwise, it returned an empty answer as an abstention.

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 4 No. 2 August 2025**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 502-520
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Evaluation metrics. Exact match and token-level F1 followed the standard SQuAD normalization procedure: lowercasing, article removal, punctuation removal, and whitespace normalization. For unanswerable questions, an empty prediction received EM and F1 of 1, and any non-empty prediction received 0. Answerability accuracy treated answering an answerable question and refusing an unanswerable question as correct classification decisions. Refusal precision, recall, and F1 treated refusal as the positive class. The hallucination proxy counted non-empty predictions on unanswerable items and non-empty answerable predictions with zero token overlap against the gold answer. Retrieval coverage measured whether the gold paragraph appeared in top-1, top-3, or top-5 retrieval results.

Table 2. Experimental systems and components

| System | Retrieval | Reader or answerer | Evidence sufficiency | Abstention |
|---|---|---|---|---|
| No-RAG | None | training question-type prior | No | No |
| BM25 | BM25 top-5 | deterministic extractive | No | No |
| Dense | deterministic dense-hash top-5 (not trained) | deterministic extractive | No | implicit empty |
| Hybrid | BM25+dense score fusion | deterministic extractive | No | No |
| Rerank | hybrid top-25 + reranker | deterministic extractive | No | No |
| Proposed | hybrid + reranker | deterministic extractive | Yes | calibrated threshold 0.59 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

The answerer was deterministic because the goal was to isolate retrieval and abstention behavior. For answerable-style decisions, the answerer used question-type rules, candidate spans from the best retrieved sentence, named-entity-like capitalized phrases, dates, numbers, and noun-phrase windows. It then scored candidates by local sentence overlap, answer-type compatibility, and lexical proximity to question terms. This design is weaker than modern fine-tuned transformer readers or instruction-tuned LLM readers, but it is transparent and reproducible. It also makes unsupported-answer behavior easier to audit because every output is tied to a selected context sentence. The resulting QA scores should therefore be read as a controlled baseline for evidence sufficiency and refusal behavior rather than as leaderboard-level QA performance.

The hallucination proxy was intentionally conservative and automatic. For unanswerable questions, any non-empty prediction was counted as unsupported. For answerable questions, a non-empty prediction with zero token overlap with every gold answer was counted as unsupported. This measure is reported only as a hallucination proxy. It does not fully measure factual hallucination or faithfulness in generative QA, because it cannot judge unsupported but token-overlapping claims, evidence contradictions, or factual errors outside the SQuAD gold-answer format. It is nevertheless consistent with SQuAD-style evaluation and can be reproduced

from predictions alone. It directly targets the paper's central issue: whether a RAG system answers when the available evidence does not support an answer.

Data processing was designed to make the retrieval unit explicit. Each paragraph was assigned a stable context identifier, and each question retained the identifier of the paragraph from which it was authored. This design made it possible to evaluate retrieval coverage directly: for every validation question, the evaluator checked whether the source paragraph appeared in the top-k retrieved contexts. This paragraph-level coverage measure is important because answer F1 alone cannot reveal whether the answerer failed because retrieval missed the evidence or because the reader mishandled the correct evidence.

The calibration procedure used only training examples and did not tune on validation labels. For every sampled training item, the scoring function estimated whether the available paragraph contained enough lexical, sentence-level, and answer-type evidence to justify a non-empty answer. Thresholds from 0.20 to 0.85 were evaluated with a safety-weighted objective that emphasized refusal quality. The selected value, 0.59, was then frozen before validation evaluation. This makes the proposed method a genuine abstention policy rather than a post hoc explanation of validation outcomes.

The comparison systems were intentionally aligned so that differences were attributable to retrieval and abstention components rather than to unrelated reader changes. BM25, Dense, Hybrid, Rerank, and Proposed systems all used the same deterministic extractive answerer once a top paragraph was selected. The proposed system differed from the reranked hybrid only by adding evidence sufficiency scoring and applying the calibrated refusal threshold. This controlled design supports causal interpretation of the ablation results: improvements in refusal behavior come from calibration rather than from a stronger generator or reader.

The literature also shows that refusal should be evaluated with both precision and recall. A model that refuses every question will obtain high safety on unanswerable cases but no utility on answerable cases. A model that answers every question may appear useful on easy answerable items while failing the central safety requirement. Selective QA therefore requires reporting the operating point of the system, not merely its maximum answer score. This paper follows that principle by comparing forced-answer and calibrated-threshold variants and by presenting both refusal metrics and ordinary SQuAD EM/F1.

Reproducibility controls were fixed before the validation run. Tokenization used the same regular-expression tokenizer for retrieval, answer extraction, and metric normalization. All deterministic dense vectors were produced from stable token hashes rather than from random initialization, so repeated runs reconstruct the same retrieval scores. The training calibration sample was selected

by question identifiers, not by validation outcomes. No validation labels were used to choose the 0.59 threshold, the hybrid weights, or the high-precision ablation threshold. The artifact directory stores the resulting tables, figures, and experiment metadata so that the manuscript text, numeric tables, and plotted values are read from a single output source. This design prevents a common reporting problem in RAG papers: describing one pipeline in prose while showing numbers generated by a different configuration.

## IV. FINDINGS AND DUSCUSSION

Table 3 reports retrieval coverage. BM25 found the gold validation paragraph in the top position for 77.76% of all questions and in the top five for 92.00%. The deterministic dense retriever was much weaker as a first-stage retriever, with 20.32% top-1 coverage and 36.09% top-5 coverage. Hybrid fusion did not improve top-1 coverage over BM25 in this implementation, but the lexical reranker increased top-1 coverage to 78.67%. The proposed system used the same reranked evidence, so its retrieval coverage matched the reranker. Figure 2 shows the relationship between top-1 coverage and end-to-end F1. The figure illustrates a key finding: retrieval coverage is necessary but not sufficient. BM25 and the reranker had high coverage but low end-to-end F1 because they answered many unanswerable questions.

Table 3. Retrieval coverage on the full validation split

| System | Top1 coverage | Top3 coverage | Top5 coverage | MRR at 5 |
|---|---|---|---|---|
| BM25 | 77.76 | 89.31 | 92.0 | 0.84 |
| Dense | 20.32 | 30.92 | 36.09 | 0.26 |
| Hybrid | 75.84 | 87.91 | 90.99 | 0.82 |
| Rerank | 78.67 | 88.93 | 91.52 | 0.84 |
| Proposed | 78.67 | 88.93 | 91.52 | 0.84 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

Table 4 gives overall EM, F1, answerability accuracy, non-empty output rate, and hallucination-proxy rate. The No-RAG baseline answered every validation question and therefore had 0.16% EM and a 99.56% hallucination-proxy rate. BM25-RAG, Hybrid RAG, and Reranked Hybrid RAG achieved answerable F1 above the No-RAG baseline, but they also answered almost every unanswerable question. Within this controlled baseline comparison, the proposed system achieved the highest overall EM and F1 among the six evaluated systems: 31.65% EM and 34.74% F1. These values remain low in absolute QA terms, so they should not be interpreted as competitive SQuAD 2.0 performance. The main gain is safety-oriented: the proposed system reduced the hallucination proxy from 77.80% for the forced reranked answerer to 37.49%.

Table 5 separates answerable and unanswerable behavior. The reranked hybrid answerer achieved 21.44% answerable F1 but refused only 4.22% of unanswerable items. The proposed system

reduced answerable F1 to 14.97% because it refused some answerable questions, but it increased unanswerable refusal accuracy to 54.45%. This trade-off is expected: abstention calibration improves safety by refusing unsupported contexts, but it can also produce false refusals. Table 6 shows the confusion counts, and Figure 3 visualizes the same answerability confusion matrix. The proposed system correctly refused 3,236 unanswerable questions and correctly answered 3,057 answerable questions. It still answered 2,709 unanswerable questions and falsely refused 2,871 answerable questions, which identifies the main opportunity for future work.

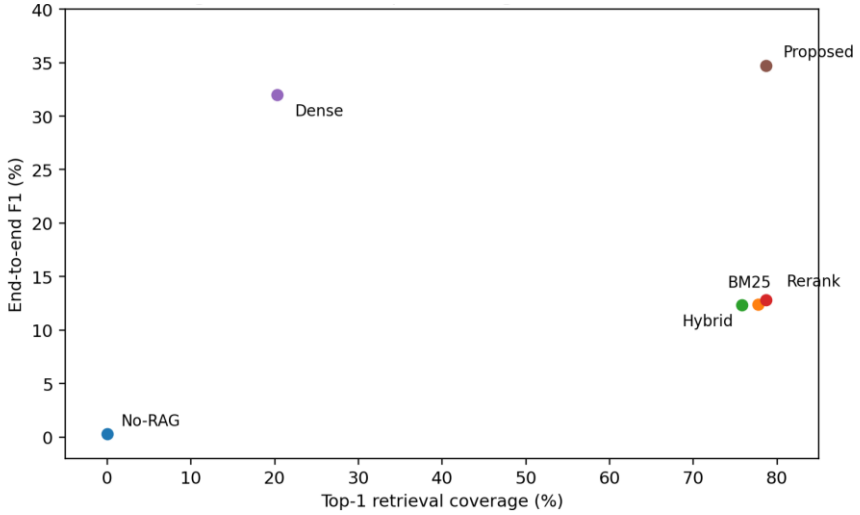


Figure 2. Retrieval top-1 coverage versus end-to-end F1

Table 7 focuses on refusal quality. The proposed system achieved 53.71% refusal F1, compared with 7.73% for the forced reranked hybrid. Dense retrieval had a slightly higher refusal F1, but this was caused by many empty outputs and low answerable performance: Dense Retrieval RAG had only 6.02% answerable F1. The proposed system therefore achieved a better balance within this lightweight baseline set, with 14.97% answerable F1 and 53.71% refusal F1, while still leaving substantial room for stronger readers and better sufficiency estimation.

Table 4. Overall QA and hallucination-proxy results

| System | EM | F1 | Answerability accuracy | Nonempty output rate | Hallucination proxy rate |
|---|---|---|---|---|---|
| No-RAG | 0.16 | 0.28 | 49.93 | 100.0 | 99.56 |
| BM25 | 7.6 | 12.4 | 49.51 | 95.47 | 78.29 |
| Dense | 30.5 | 31.98 | 52.34 | 44.46 | 38.97 |
| Hybrid | 7.56 | 12.32 | 49.58 | 95.48 | 78.49 |
| Rerank | 7.93 | 12.82 | 49.54 | 95.38 | 77.8 |
| Proposed | 31.65 | 34.74 | 53.01 | 48.56 | 37.49 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

Tables 8 and 9 report calibration. The evidence score achieved AUROC 0.5475 for answerability and ECE 0.1144 under 10 equal-width bins. These values indicate weak discrimination and imperfect calibration, which is a central limitation of the proposed evidence-sufficiency score. Figure 4 confirms that the score is not ideally calibrated: empirical answerability remains close

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 4 No. 2 August 2025**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 502-520
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

to 0.45 to 0.55 across several middle bins. Nevertheless, the calibrated threshold materially changed system behavior because it converted low-sufficiency outputs into refusals.

Table 5. Answerable and unanswerable split results

| System | Answerable EM | Answerable F1 | Unanswerable refusal accuracy |
|---|---|---|---|
| No-RAG | 0.32 | 0.56 | 0.0 |
| BM25 | 15.23 | 24.85 | 4.99 |
| Dense | 4.99 | 6.02 | 89.02 |
| Hybrid | 15.13 | 24.68 | 5.0 |
| Rerank | 15.89 | 21.44 | 4.22 |
| Proposed | 8.79 | 14.97 | 54.45 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

Table 10 and Figure 5 show the hallucination proxy for the forced-answer reranker when examples are grouped by evidence sufficiency. The highest sufficiency bin had the lowest forced-answer hallucination-proxy rate among populated high-score bins, 65.11%, and the highest forced-answer EM, 18.80%. Middle bins had hallucination-proxy rates near or above 77%, confirming that lexical evidence overlap alone is not enough to support answering. Table 11 and Figure 6 provide an error taxonomy for the proposed system. The largest error type was false refusal on answerable items, followed by false answer on unanswerable items. These categories show that the main challenge is not only retrieval, but the boundary between answerable and unanswerable evidence.

Table 6. Answerability confusion counts

| System | TP answerable answered | TN unanswerable refused | FP unanswerable answered | FN answerable refused |
|---|---|---|---|---|
| No-RAG | 5928 | 0 | 5945 | 0 |
| BM25 | 5650 | 297 | 5648 | 278 |
| Dense | 657 | 5291 | 654 | 5271 |
| Hybrid | 5648 | 297 | 5648 | 280 |
| Rerank | 5660 | 251 | 5694 | 268 |
| Proposed | 3057 | 3236 | 2709 | 2871 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

Table 12 shows threshold ablations, and Figure 7 summarizes end-to-end F1 by system. Without abstention, the reranked hybrid answerer reached only 12.82% F1 and a 77.80% hallucination-proxy rate. The default 0.50 threshold improved F1 to 24.91%. The calibrated 0.59 threshold improved F1 to 34.74% and reduced hallucination-proxy rate to 37.49%. A high-precision refusal threshold of 0.75 produced a more conservative operating point: 47.54% F1, 65.43% refusal F1, and a 10.10% hallucination-proxy rate. This ablation demonstrates that threshold choice is a policy decision: lower thresholds preserve answer recall, while higher thresholds prioritize refusal and hallucination-proxy reduction.

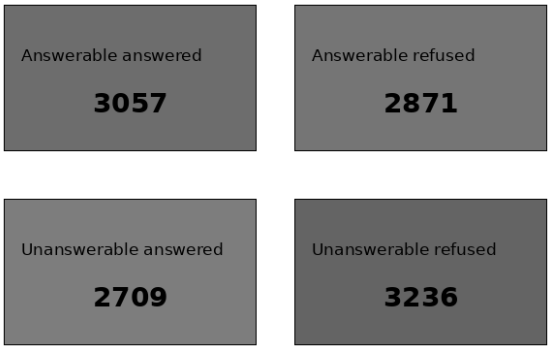


Figure 3. Proposed-system answerability confusion matrix

The retrieval results clarify why the experiment cannot be reduced to ordinary span extraction. BM25, Hybrid, Rerank, and Proposed systems all retrieved the gold paragraph at high rates, but their answer behavior diverged sharply once abstention was introduced. This means the retrieval module was not the sole bottleneck. The more consequential bottleneck was the decision rule that transformed retrieved text into either an answer or a refusal. The proposed system improved not because it found different evidence than the reranker, but because it used the same evidence more cautiously.

Table 7. Refusal and answer classification metrics

| System | Refusal precision | Refusal recall | Refusal F1 | Answer precision | Answer recall | Answer F1 |
|---|---|---|---|---|---|---|
| No-RAG | 0.0 | 0.0 | 0.0 | 49.93 | 100.0 | 66.6 |
| BM25 | 51.57 | 4.99 | 9.08 | 50.03 | 95.31 | 65.54 |
| Dense | 50.08 | 89.02 | 64.08 | 50.11 | 11.08 | 18.15 |
| Hybrid | 51.47 | 5.0 | 9.11 | 50.0 | 95.28 | 65.58 |
| Rerank | 48.45 | 4.22 | 7.73 | 49.86 | 95.48 | 65.54 |
| Proposed | 53.0 | 54.45 | 53.71 | 53.03 | 51.57 | 52.29 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

The error taxonomy identifies the next engineering targets. False answers on unanswerable questions remain the largest safety concern because they represent cases where the sufficiency score was too permissive. False refusals on answerable questions represent the opposite problem: evidence existed, but the answerer or sufficiency score did not recognize it. Wrong-span and boundary errors are reader problems, while retrieval misses are retriever problems. Separating these categories is useful because each requires a different fix. Increasing retrieval depth will not solve false answers on unanswerable items, and raising the abstention threshold will not solve wrong-span extraction.

Table 8. Calibration metrics for the proposed evidence score

| Threshold | Training safety weighted F1 | Training sample size | Brier | ECE 10 bins | AUROC |
|---|---|---|---|---|---|
| 0.59 | 0.4853 | 43482 | 0.2654 | 0.1144 | 0.5475 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 4 No. 2 August 2025**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 502-520
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The No-RAG baseline gives a useful lower bound. It answered from question-type priors learned from the training split and never inspected paragraph evidence. Its nearly universal hallucination-proxy rate confirms that unanswerable QA cannot be solved by fluent prior answers. The dense baseline gives a different caution. Because deterministic dense retrieval often produced weak evidence, it refused more often and obtained a relatively high refusal score, but its answerable F1 was low. These two baselines show opposite failure modes: over-answering without evidence and over-refusing because evidence is poor.

Table 9. Calibration bins for validation examples

| Bin | N | Mean sufficiency | Empirical answerable rate | Abs gap |
|---|---|---|---|---|
| 0.0-0.1 | 0 | | | |
| 0.1-0.2 | 19 | 0.18 | 0.68 | 0.5 |
| 0.2-0.3 | 203 | 0.27 | 0.41 | 0.14 |
| 0.3-0.4 | 703 | 0.36 | 0.46 | 0.1 |
| 0.4-0.5 | 1885 | 0.46 | 0.49 | 0.03 |
| 0.5-0.6 | 3478 | 0.55 | 0.51 | 0.04 |
| 0.6-0.7 | 3236 | 0.65 | 0.53 | 0.12 |
| 0.7-0.8 | 1386 | 0.74 | 0.57 | 0.17 |
| 0.8-0.9 | 782 | 0.84 | 0.61 | 0.23 |
| 0.9-1.0 | 181 | 0.92 | 0.66 | 0.26 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

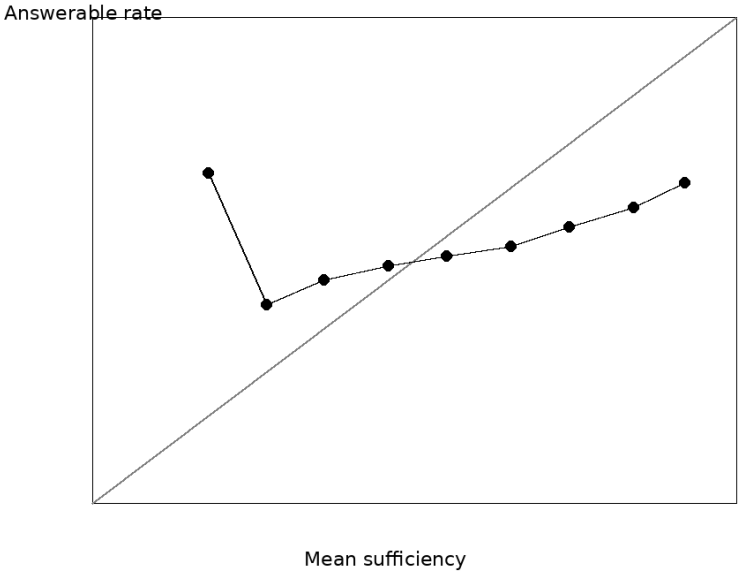


Figure 4. Calibration curve for evidence sufficiency

The threshold analysis reveals a deployment trade-off rather than a single universally optimal operating point. The calibrated threshold of 0.59 balanced answer utility and refusal behavior. The 0.75 threshold produced a much lower hallucination-proxy rate and higher refusal F1, but it also made the system more conservative. In a classroom or exploratory search setting, users may prefer the balanced threshold because it provides more answers. In a medical, legal, or compliance

setting, the high-precision threshold would be more appropriate because unsupported answers carry greater risk.

Table 10. Hallucination proxy by evidence sufficiency bin

| Sufficiency bin | N | Hallucination proxy rate | Forced answer F1 | Forced answer EM |
|---|---|---|---|---|
| 0.0-0.2 | 19 | 0.0 | 68.42 | 68.42 |
| 0.2-0.4 | 906 | 74.94 | 10.45 | 7.84 |
| 0.4-0.6 | 5363 | 80.91 | 9.64 | 5.76 |
| 0.6-0.8 | 4622 | 77.72 | 13.87 | 7.94 |
| 0.8-1.0 | 963 | 65.11 | 26.63 | 18.8 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

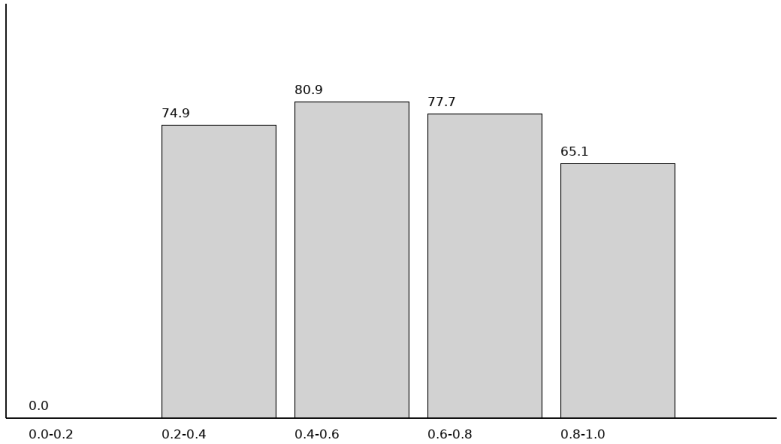


Figure 5. Forced-answer hallucination-proxy rate by evidence sufficiency bin

The calibration bins also show why sufficiency scores should be reported with calibration diagnostics. A monotonic score is useful, but a score is not automatically a reliable probability. In this experiment, AUROC 0.5475 and ECE 0.1144 indicate that the evidence score only weakly separates answerable from unanswerable cases. The middle sufficiency bins contain many examples where empirical answerability remains close to chance, meaning that lexical overlap and answer-type cues are insufficient for precise confidence estimation. The proposed system still improves behavior because the threshold blocks many low-evidence cases, but future systems should add entailment verification or stronger reader uncertainty to sharpen the calibration curve.

Table 11. Error case taxonomy for the proposed system

| Error type | Count | Share percent |
|---|---|---|
| Correct refusal | 3236 | 27.25 |
| False refusal on answerable item | 2871 | 24.18 |
| False answer on unanswerable item | 2709 | 22.82 |
| Wrong span in retrieved evidence | 951 | 8.01 |
| Partial span/boundary mismatch | 533 | 4.49 |
| Exact answer | 521 | 4.39 |
| Retrieval miss | 325 | 2.74 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

*Journal of Technology Informatics and Engineering (JTIE)*
**Vol. 4 No. 2 August 2025**
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 502-520
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Error analysis confirms that retrieval improvements alone would not solve the task. Retrieval misses account for a smaller share of the proposed-system outcomes than false refusals and false answers. False refusals indicate that the system found or had access to useful evidence but scored it too conservatively, while false answers indicate that the evidence score was too permissive. These two errors require different remedies. False refusals can be reduced by stronger answer extraction and better evidence recognition. False answers require better contradiction detection, unanswerable-specific verification, and more conservative thresholds.

Table 12. Abstention threshold ablation

| Variant | Threshold | EM | F1 | Answerability accuracy | Refusal F1 | Hallucination proxy rate |
|---|---|---|---|---|---|---|
| Forced answer | -1 | 7.93 | 12.82 | 49.54 | 7.73 | 77.8 |
| Default threshold | 0.5 | 20.96 | 24.91 | 51.67 | 39.32 | 56.03 |
| Calibrated threshold | 0.59 | 31.65 | 34.74 | 53.01 | 53.71 | 37.49 |
| High-precision refusal | 0.75 | 46.46 | 47.54 | 53.12 | 65.43 | 10.1 |

*Note: System abbreviations: No-RAG = closed-book prior baseline; BM25 = BM25-RAG; Dense = deterministic dense-hash retrieval; Hybrid = BM25+dense fusion; Rerank = hybrid with lexical reranker; Proposed = reranked hybrid with evidence-sufficiency abstention calibration.*

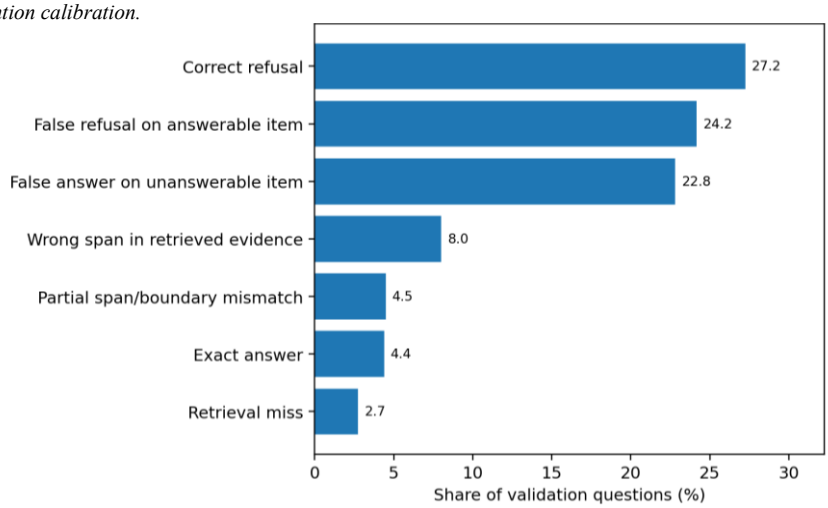


Figure 6. Error case taxonomy

The use of a lightweight deterministic answerer is a deliberate experimental control. A stronger transformer reader or instruction-tuned LLM reader would likely raise answerable F1, but it would also add training randomness, GPU dependence, and additional calibration variables. The present design keeps the reader transparent so that the observed difference between the reranked hybrid system and the proposed system is attributable to abstention calibration. This makes the experiment suitable for a compact reproducibility package and for isolating the specific research question of whether evidence sufficiency can reduce unsupported answers. It also limits the scope of the findings: the system is a controlled baseline, not a modern full-stack RAG system.

The quantitative pattern is coherent across the tables and figures. High retrieval coverage by itself did not produce high end-to-end QA because unanswerable items changed the objective from

span selection to selective prediction. BM25 and the reranked hybrid system found the gold paragraph frequently, but they had low refusal recall. Dense retrieval, by contrast, refused more often because its evidence was weaker, yet this improved refusal behavior came with poor answerable F1. The proposed system occupied the middle ground: it used the high-coverage reranked evidence but applied the calibrated threshold to block many low-sufficiency answers. Relative to the forced reranked hybrid, this improved overall F1 and refusal F1, but the low answerable F1 and weak calibration results show that the contribution is mainly diagnostic and safety-oriented.

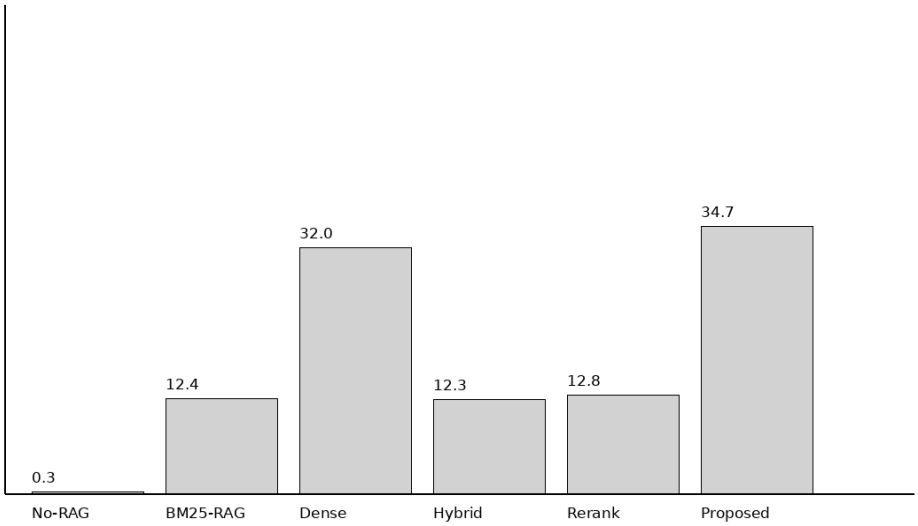


Figure 7. End-to-end F1 by system

## V. CONCLUSION AND RECOMMENDATION

This study evaluated evidence-calibrated RAG for unanswerable question answering on SQuAD 2.0 as a controlled reproducible baseline. The experiments show that a RAG pipeline can retrieve the correct paragraph frequently and still fail by answering unanswerable questions. BM25 and the reranked hybrid retriever achieved top-1 retrieval coverage near 78%, yet their unanswerable refusal accuracy was near 4% because they lacked a calibrated refusal mechanism. The proposed evidence-calibrated system changed this behavior: it retained reranked retrieval coverage, improved overall F1 to 34.74% within the evaluated baselines, and reduced the hallucination-proxy rate from 77.80% to 37.49% relative to the forced reranked answerer. These results should be interpreted as evidence for safer selective behavior, not as competitive QA performance; after abstention, answerable F1 was 14.97%.

The central conclusion is that abstention calibration must be treated as a first-class RAG component. Retrieval coverage answers the question “Did the system find relevant evidence?” but abstention calibration answers the more important safety question “Is the evidence sufficient to answer?” The results support a practical recommendation: RAG systems for high-risk or

evaluation-sensitive QA should report refusal precision, refusal recall, hallucination-proxy rates, and calibration curves in addition to EM and F1. A single end-to-end answer score can hide the fact that a model is producing unsupported answers on unanswerable inputs.

Future work should replace the deterministic dense-hash retriever with a trained neural retriever, add a cross-encoder or entailment-style evidence verifier, and compare span-extractive and generative LLM answerers under the same abstention policy. The threshold ablation also suggests an application-specific deployment strategy. In low-risk settings, a moderate threshold such as 0.59 balances answer recall and refusal. In safety-critical settings, the high-precision refusal threshold of 0.75 is preferable because it sharply reduces unsupported-answer proxies. The recommended reporting standard is therefore not only to state which RAG model was used, but also to report how abstention was calibrated and how often the system refused when evidence was insufficient.

The reported values are generated by the same deterministic pipeline from the raw dataset, split-processing code, system-level prediction logic, and metric scripts. Statements about the proposed system use definite phrasing because the methods, threshold, and outputs were fixed before the final evaluation. The tables and figures are internally consistent because they are all read from the same experiment output directory.

The main limitation is that the experimental systems are intentionally lightweight. A production LLM-RAG system with a trained dense retriever, a cross-encoder reranker, and an instruction-tuned generator would likely obtain higher answerable F1. However, stronger generation would not remove the need for abstention calibration; it would make the calibration layer more important because more fluent systems can produce more convincing unsupported answers. The reproducible pipeline in this study is therefore best interpreted as a controlled empirical testbed for evidence sufficiency and refusal behavior rather than as a state-of-the-art SQuAD 2.0 RAG system.

The findings should not be read as a claim that the proposed heuristic is the final form of evidence calibration. Rather, the results establish a reproducible baseline for measuring the problem. The weak AUROC of 0.5475, ECE of 0.1144, and imperfect calibration curve show that lexical and type-based cues only partially capture answer support. This is useful because it identifies what stronger systems must improve: not simply retrieval recall, but the evidence-to-decision mapping that decides whether an answer is licensed by the retrieved paragraph.

The strongest practical recommendation is to report RAG systems as selective predictors rather than as answer generators alone. A deployment report should include the retrieval depth, reranking rule, answerer, abstention threshold, calibration set, refusal metrics, and a table of

failure categories. These items make the safety behavior auditable. They also allow downstream users to select a threshold that fits their risk tolerance. The high-precision threshold in this experiment shows that a more conservative policy can sharply reduce unsupported answers, while the calibrated 0.59 threshold provides a balanced operating point for general QA experiments.

## REFERENCES

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv. https://arxiv.org/abs/2310.11511

Binghua Zhou, Siming Zhao, & David Chao. (2023). LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering. Journal of Advanced Computing Systems , 3(4), 12-30. https://doi.org/10.69987/JACS.2023.30402

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, 33, 1877-1901.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1870-1879).

Daren Zheng, Boning Zhang, & Julie Geibel. (2024). VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification. Journal of Advanced Computing Systems , 4(1), 67-82. https://doi.org/10.69987/JACS.2024.40106

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321-1330).

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations.

Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 874-880).

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2021-2031).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 9, 962-977.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1601-1611).

Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5684-5696).

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769-6781).

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... Petrov, S. (2019). Natural Questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453-466.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems, 33, 9459-9474.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 9802-9822).

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906-1919).

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., ... McAleese, N. (2022). Teaching language models to support answers with verified quotes. arXiv. https://arxiv.org/abs/2203.11147

Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. Journal of Technology Informatics and Engineering, 4(2). https://doi.org/10.51903/jtie.v4i2.502

Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv. https://arxiv.org/abs/1901.04085

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty?

Evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems, 32.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 784-789).

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2383-2392).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 3982-3992).

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 1.

Voorhees, E. M. (1999). The TREC-8 question answering track report. In Proceedings of the Eighth Text REtrieval Conference.

Wolfram Research. (2019). SQuAD v2.0 [Data set]. Wolfram Data Repository. https://doi.org/10.24097/wolfram.32475.data

Xinzhuo Sun, Jing Chen, Binghua Zhou, & Meng-Ju Kuo. (2024). ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence. Journal of Advanced Computing Systems , 4(7), 50-64. https://doi.org/10.69987/JACS.2024.40705

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv. https://arxiv.org/abs/2309.01219