CITADEL — Citation-Driven Draft-Evaluate Loop

Daniel Seredensky, Dylan Iddings, Sharon G. Small
The Siena University Institute for Artificial Intelligence
Siena University, Loudonville, NY 12211
{dp17sere, dt06iddi, ssmall}@siena.edu

Abstract

This paper describes our team submission from the Siena University Institute for Artificial Intelligence for the Text Retrieval Conference (TREC) 2025 Detection, Retrieval, and Augmented Generation for Understanding News Track (DRAGUN). Our approach combines classical retrieval methods with the usage of multiple large language model (LLM) agents to generate concise, evidence-based reports. First, a hybrid retrieval pipeline integrates BM25, synonym-based query expansion, and cross-encoder reranking to maximize recall and precision across the corpus. Then, retrieved passages are processed through a generation-evaluation loop, where different LLM agents separately generate and critique reports according to rubric-based criteria for coverage, accuracy, and citation quality. This design emphasizes factual accuracy and access to citations to align with DRAGUN's goal of supporting critical engagement with news.

1 Introduction

The purpose of the DRAGUN track is to support readers in evaluating the trustworthiness of online news by guiding participants through two complementary tasks. Task 1 consists of generating critical questions that a reader should ask about a news article. Task 2 is focused producing an evidence-based report that addresses these questions. The goal is to provide contextual information, analyze sources, and present multiple perspectives to help readers form their own judgments without asserting definitive conclusions. (Zhang et al., 2025)

Our team decided to focus our efforts on Task 2: generating an evidence-based report grounded in relevant source documents. The system begins with a hybrid retrieval strategy that combines BM25 scoring, synonym-based query expansion, and cross-encoder reranking to locate high-quality evidence from the corpus. Retrieved passages are then processed through an iterative generation–evaluation loop, in which LLM agents refine a draft report through rubric-based critique and revision. The final output is a structured, citation-supported document that assists readers in forming their own judgments about an article.

2 The DRAGUN Track

The Detection, Retrieval, and Augmented Generation for Understanding News (DRAGUN) Track at TREC 2025 builds on the 2024 Lateral Reading Track (Zhang et al., 2024), extending its goal of supporting the evaluation of online news trustworthiness. While the previous track focused on how systems could retrieve evidence to help readers assess information, DRAGUN shifts attention to how systems can use that evidence to generate complete reports. These reports are designed to help readers understand whether a news article is credible, biased, or missing important context.

In DRAGUN, each topic is a real-world news article. For every topic, participating systems must create a 250-word report that draws upon evidence from the MS MARCO V2.1 (Segmented) corpus, which contains over 114 million passages. Each sentence in the report can cite up to three passages, allowing detailed reference and source verification. This design connects each claim in the generated text to specific retrieved evidence, making it possible to trace how conclusions were reached.

Instead of evaluating retrieval results directly, NIST assessors create investigative questions about each news article and then judge whether a system's report answers them accurately and completely. This evaluation method replaces ranking-based retrieval measures with an assessment of how informative, correct, and well-supported each report is (Zhang et al., 2024). It tests whether systems can produce useful summaries that help readers make informed judgments about online news.

The shift from retrieval to generation means that DRAGUN now measures how well systems can combine information and explain it, rather than only how well they can find relevant evidence. DRAGUN asks systems not only to identify supporting documents but also to combine them into clear and verifiable text. By focusing on how information is summarized, cited, and reasoned through, the track aims to measure real progress in retrieval-augmented generation methods for news understanding.

Overall, DRAGUN presents a unified framework for examining how automated systems can support critical reading at scale. By integrating retrieval, synthesis, and reporting into a single task, it provides a practical way to assess whether language models can produce grounded, transparent, and trustworthy summaries of real-world news.

3 Implementation Details

3.1 Design Principles

We followed two main ideas when building CITADEL. First, we used a recall-first, precision-later pipeline. In NLP systems for fact-checking—especially scientific verification—this balances precision and recall: early stages gather as much potentially relevant material as possible; later stages refine by ranking and filtering (Glass et al., 2022). Second, we applied separation of concerns across LLM roles to better control the context window and to minimize the effect of the model "grading its own paper," a known issue in LLM-as-judge setups (Wataoka et al., 2024; Li et al., 2024).

3.2 Retrieval Layer

At the root of the system is information retrieval. We used a Lucene index of the MS MARCO v2.1 segmented corpus, distributed via Pyserini and hosted by the University of Waterloo's RGW storage service (Nguyen et al., 2016; Lin et al., 2024). Our system's Java service exposes two methods—SelectDocuments and Search. SelectDocuments returns full passages by segment ID. The process runs a reader thread (stdin), a writer thread (stdout), and a fixed pool of worker threads for handling queries. Search accepts a main query plus several sub-queries; each sub-query is expanded with synonyms from a prebuilt WordNet map (Princeton University, 2010) and executed in parallel using BM25 as the ranking algorithm. We merge results at the full-document level using Reciprocal Rank Fusion (RRF). WordNet gives a structured, interpretable synonym set that mitigates vocabulary mismatch—classic query expansion motivation (Miller, 1995). RRF reliably improves performance when combining rankings (e.g., multiple query variants) without heavy tuning (Cormack et al., 2009). BM25 is our first-stage ranker for its established, strong performance and probabilistic grounding; we did not tune beyond standard practice (Robertson and Zaragoza, 2009).

A lightweight Python client communicates with the Java service using framed JSON over stdin/stdout. The client takes the ranked results and sends the top-200 hits (segments collapsed on full document id) to a Cohere cross-encoder re-ranker using the master query. Cross-encoders are well known to boost quality when re-ranking a fast lexical candidate set; multi-stage designs limit expensive inference to a small shortlist (Nogueira and Cho, 2019). After re-ranking, we forward only the top-15 metadata entries (URL, title, authors, headers, etc) to the agentic layer for document selection. This keeps downstream prompts compact while retaining the most useful evidence.

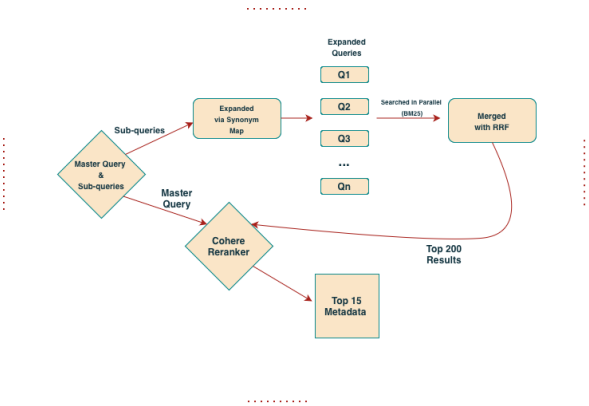

Figure 1: End-to-end IR Module

3.3 IR Agents

The Python client's search and select methods are exposed as tools to a lightweight IR-Agent (GPT-4.1-mini). The agent runs a gather-information → update-answer loop over question–document-context pairs. This can be seen in the example below.

Excerpt from Appendix A, Listing 8 (IR-Agent Search (CoT)):

```
<cot>
For each question, I will generate up to 4 BM25-
    optimized keyword queries targeting specific
     aspects, and a master query for semantic
    reranking. The queries will focus on public/
```

```
    media reaction and restaurant marketing/
    serving methods. I reaffirm that my answer
    will be valid JSON.
</cot>
```

Excerpt from Appendix A, Listing 9 (IR-Agent Search (Answer payload)):

```
<answer>
{
   "searches": [
      {
         "queries": [
            "initial public reaction to
                Hawaiian pizza invention",
            "media coverage of Hawaiian pizza
                after it was created",
            "controversy surrounding Hawaiian
                pizza in its early years",
            "reception of pineapple pizza by
                public and press"
         ],
         "master_query": "What was the initial
             public and media reaction to
            Hawaiian pizza after its
            invention, including any
            controversy or divisiveness?"
      },
      {
         "queries": [
            "how Satellite restaurant marketed
                Hawaiian pizza",
            "early advertising strategies for
                Hawaiian pizza by Satellite
                restaurant",
            "methods used to serve Hawaiian
                pizza at Satellite restaurant
                ",
            "Panopoulos promotion of pineapple
                pizza in the restaurant's
                early years"
         ],
         "master_query": "How did the
             Satellite restaurant market and
            serve Hawaiian pizza in its early
             years, including advertising and
             serving methods?"
      }
   ]
}
</answer>
```

Excerpt from Appendix A, Listing 10 (IR-Agent Select (CoT)):

```
<cot>
I reviewed the search metadata for relevance to
    the questions about initial public/media
    reaction to Hawaiian pizza and how the
    Satellite restaurant marketed and served it.
     I selected segments that likely contain
    historical context, controversy, marketing,
    and inventor details. I reaffirm that my
    answer will be valid JSON.
</cot>
```

Excerpt from Appendix A, Listing 11 (IR-Agent Select (Answer payload)):

```
<answer>
{
  "selections": [
    "msmarco_v2.1_doc_28_903296016#2_1908541802",
    "msmarco_v2.1_doc_30_523003770#0_1183664316",
    "msmarco_v2.1_doc_57_619272585#9_900954951",
    "msmarco_v2.1_doc_51_759827373#0_1589150899",
    "msmarco_v2.1_doc_16_565495028#0_805615245",
    "msmarco_v2.1_doc_17_1448273400#0_1598351114"
  ]
}
</answer>
```

Excerpt from Appendix A, Listing 12 (IR-Agent Update (CoT)):

```
<cot>
Using the provided segments, I can confirm that
    the initial public and media reaction to
    Hawaiian pizza was divisive and
    controversial, but detailed early reactions
    are not extensively documented. The pizza
    was invented by Sam Panopoulos in 1962 at
    the Satellite Restaurant in Ontario, Canada,
     where he and his brothers experimented with
     pineapple on pizza to attract customers.
    The pizza was named after the brand of
    canned pineapple used. Early popularity was
    limited but grew locally and then globally.
    The Satellite restaurant marketed it by
    experimenting with new flavors, but specific
     marketing strategies are not detailed. I
    reaffirm that my answer will be valid JSON.
</cot>
```

Excerpt from Appendix A, Listing 13 (IR-Agent Update (Answers + rounds)):

```
<answer>
{
  "questions": [
    {
      "question": "What was the initial public
          and media reaction to Hawaiian pizza
          after its invention?",
      "doc_context": "The article mentions
          controversy and divisiveness but does
          not detail early reactions.",
      "answer": {
        "text": "The initial public and media
            reaction to Hawaiian pizza was
            divisive and controversial, with the
            pizza inspiring jokes, memes, and
            debates. However, detailed early
            reactions are not extensively
            documented. The pizza was invented in
             1962 by Sam Panopoulos, who defended
             his creation throughout his life.
            While the addition of pineapple was
            not initially very popular, it
            eventually became a staple offering.
            The controversy mainly centers on the
             inclusion of pineapple as a topping
            .",
        "citations": [
          {
            "summary": "Describes the
                divisiveness and controversy
```

```
            surrounding Hawaiian pizza and
            its inventor Sam Panopoulos.",
          "citation": "msmarco_v2.1
            _doc_28_903296016#2_1908541802"
        },
        {
          "summary": "Details the invention of
            Hawaiian pizza by Sam Panopoulos
            in 1962 and the naming after the
            canned pineapple brand.",
          "citation": "msmarco_v2.1
            _doc_16_565495028#0_805615245"
        },
        {
          "summary": "Notes the controversy and
             polarizing debate mainly focused
             on pineapple as a topping.",
          "citation": "msmarco_v2.1
            _doc_17_1448273400#0_1598351114"
        }
      ]
    },
    "finished": true
  },
  {
    "question": "How did the Satellite
        restaurant market and serve Hawaiian
        pizza in its early years?",
    "doc_context": "Panopoulos began
        advertising his new pizza topping, but
         details are sparse.",
    "answer": {
      "text": "The Satellite restaurant, where
          Hawaiian pizza was invented by Sam
          Panopoulos in 1962, marketed the
          pizza by experimenting with new
          flavor combinations to attract
          customers. Panopoulos and his
          brothers added pineapple and ham to a
           standard cheese pizza as a fun
          experiment. The pizza was named after
           the canned pineapple brand used.
          Specific marketing strategies or
          serving methods beyond this
          experimentation are not detailed in
          the available information.",
      "citations": [
        {
          "summary": "Describes Panopoulos'
            experimentation with pineapple
            and ham at the Satellite
            restaurant to attract customers
            .",
          "citation": "msmarco_v2.1
            _doc_30_523003770#0_1183664316"
        },
        {
          "summary": "Notes the invention of
            Hawaiian pizza at the Satellite
            Diner and its naming after the
            canned pineapple brand.",
          "citation": "msmarco_v2.1
            _doc_57_619272585#9_900954951"
        },
        {
          "summary": "Mentions Panopoulos'
            experience and experimentation at
             the Satellite restaurant in
            Ontario, Canada.",
          "citation": "msmarco_v2.1
            _doc_16_565495028#0_805615245"
        }
      ]
    },
    "finished": true
  }
],
"rounds": [
  {
    "summary": "Initial round focused on
        queries about public/media reaction
        and Satellite restaurant marketing.
        Found relevant segments confirming
        invention details, divisiveness, and
        experimentation but no detailed early
        marketing strategies or media
        reactions.",
    "seen_ids": [
      "msmarco_v2.1_doc_28_903296016#2
          _1908541802",
      "msmarco_v2.1_doc_30_523003770#0
          _1183664316",
      "msmarco_v2.1_doc_57_619272585#9
          _900954951",
      "msmarco_v2.1_doc_16_565495028#0
          _805615245",
      "msmarco_v2.1_doc_17_1448273400#0
          _1598351114"
    ]
  }
]
}
</answer>
```


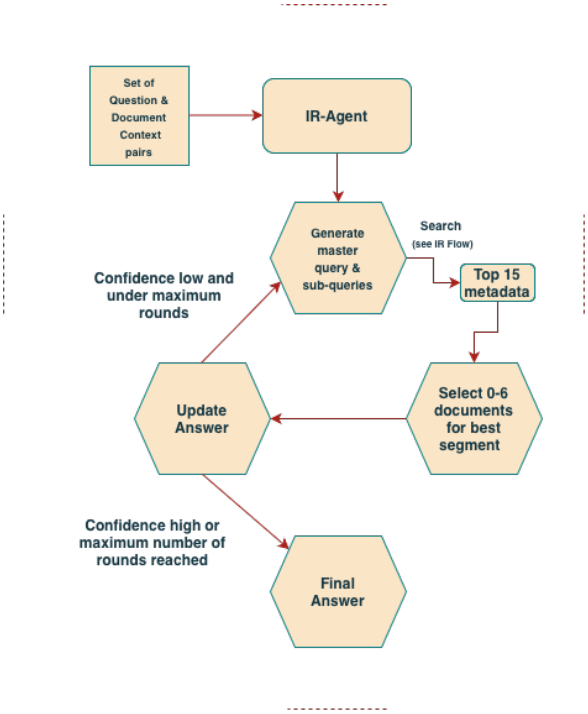

Figure 2: IR Agent Flow

3.4 Draft-Critique Loop

Above the IR layer, we split writing and critique into two distinct agents: a Report Generator and

a Report Evaluator (both GPT-4.1). In round one, the Generator produces an informational, fact-checked summary and includes a short note to the Evaluator describing missing information or desired improvements. The Evaluator reviews the same sources and the draft, returning a strict JSON rubric with the following criteria graded on a scale of 1–5:

* Coverage
* Accuracy
* Citation Quality
* Style
* Prioritization
* Completeness
* Information Density

Excerpt from Appendix A, Listing 3 (Generator–0 Report):

```
<report>
{
    "responses": [
        {
            "text": "Hawaiian pizza was invented
                in 1962 by Sam Panopoulos, a
                Greek-Canadian restaurateur in
                Ontario, Canada. Despite its name
                , the pizza has no direct
                connection to Hawai'i and was
                created in a diner serving
                typical mid-century food. The
                dish combines canned pineapple
                and ham atop a cheese pizza, and
                its creation was influenced by
                the popularity of canned
                pineapple and tiki culture in
                North America during the 1960s.",
            "citations": []
        },
        {
            "text": "Panopoulos began
                experimenting with pizza after
                observing its preparation in
                Windsor, Canada, and Detroit, USA.
                 At the time, pizza was a
                relatively new and unfamiliar
                dish in Canada, with limited
                topping options and primitive
                cooking methods. The addition of
                pineapple was inspired by the
                sweet and savory flavors found in
                 Chinese cuisine, which
                Panopoulos already served at his
                restaurant.",
            "citations": []
        },
        {
            "text": "Since its inception,
                Hawaiian pizza has become a
                divisive and controversial item
                in the pizza world, often
                polarizing opinions among
                consumers. While the classic
                combination of ham and pineapple
                was initially accidental,
                Panopoulos named the creation '
                Hawaiian pizza' from the start,
                and it quickly gained popularity
                despite its unconventional
                ingredients.",
            "citations": []
        }
    ]
}
</report>
```

Excerpt from Appendix A, Listing 7 (Evaluator–0 Rubric Scores):

```
<eval>
{
  "coverage": 4,
  "accuracy": 5,
  "citation_quality": 2,
  "style": 4,
  "prioritization": 4,
  "completeness": 3,
  "information_density": 3,
  "comments": "The report accurately summarizes
      the invention and cultural context of
      Hawaiian pizza, prioritizing key facts.
      However, it lacks granular details and
      direct quotes that would increase
      information density and completeness.
      Citation quality is low due to absence of
      references, though this is understandable
      for the first round. Coverage is strong
      but could be improved with more specifics
      about Panopoulos and the Satellite
      restaurant. Style is clear and concise."
}
</eval>
```

Although each criterion is graded 1-5, the internal evaluation heuristic places 3× weight on coverage and 2× weight on accuracy and citation quality. The Evaluator also proposes targeted follow-up questions tied to the current document context and sends a brief note back to the Generator. *Excerpt from Appendix A, Listing 6 (Evaluator–0 Questions):*

```
<ir>
{
  "questions": [
    {
      "question": "What was the initial public
          and media reaction to Hawaiian pizza
          after its invention?",
      "context": "The article mentions
          controversy and divisiveness but does
          not detail early reactions."
    },
    {
      "question": "How did the Satellite
          restaurant market and serve Hawaiian
          pizza in its early years?",
      "context": "Panopoulos began advertising
          his new pizza topping, but details are
           sparse."
    },
```

```
    {
      "question": "Are there any records or
          interviews with customers or staff
          from the Satellite restaurant
          regarding the invention?",
      "context": "The article relies on
          Panopoulos's recollections; external
          perspectives are missing."
    },
    {
      "question": "How did the recipe for
          Hawaiian pizza evolve after its
          initial creation?",
      "context": "It is noted that the ham and
          pineapple combination was accidental,
          but further evolution is not described
          ."
    }
  ]
}
</ir>
```


These questions are distributed to multiple IR-Agents running concurrently; their findings are appended to an expanding context document. *Excerpt from Appendix A, Listing 14 (Context Q1)*[1]*:*


```
{
"question": "Are there any records or interviews
     with customers or staff from the Satellite
    restaurant regarding the invention?",
"doc_context": "The article relies on Panopoulos
    's recollections; external perspectives are
    missing.", "answer": {
   "text": "There are no known records or
       interviews with customers or staff from
       the Satellite restaurant regarding the
       invention of Hawaiian pizza. The
       information primarily relies on Sam
       Panopoulos's own recollections and
       statements.",
   "citations": [
      {
      "summary": "Panopoulos created Hawaiian
          pizza at the Satellite Restaurant in
          1962, but external perspectives are
          missing.", "citation": "msmarco_v2.1
          _doc_04_420132660#5_965977302"
      },
      {
      "summary": "Panopoulos described the
          invention and experimentation with
          pineapple on pizza, with no mention
          of other interviews.",
      "citation": "msmarco_v2.1
          _doc_30_523003770#0_1183664316"
      }
   ]},
"finished": true
}
```


Each subsequent round proceeds with the accumulated notes and context available to both roles. The loop stops when the rubric scores are perfect or a round limit is reached. While tailored to this task, the general idea—iterative critique and self-correction—follows recent LLM work such as Self-Refine, Reflection, and deliberate search/planning methods; Google's "Deep Researcher with Test-Time Diffusion" describes a related draft-and-revise process (Madaan et al., 2023).

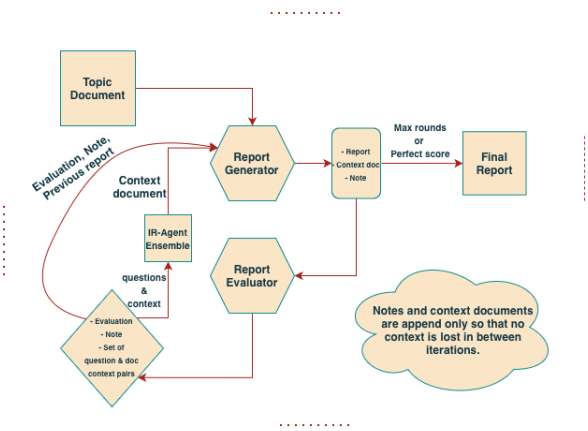

Figure 3

3.5 Systems Engineering

The Java retrieval service is started once and kept resident to avoid JVM cold starts. Python 3.11 coordinates I/O-bound tasks with asyncio and uvloop. Lucene queries run on the Java worker pool; reranking and LLM calls are handled asynchronously on the Python side. To stay within Azure and Cohere quotas, we implemented a sliding-window token bucket with unique IDs per reservation, allowing precise refunds of overestimated usage. Throughput is continuously monitored and logged to a dedicated data directory. This achieves high concurrency without breaching rate limits while keeping latency manageable.

3.6 Rationale

The pipeline follows a well-established, evidence-based pattern that balances recall, precision, and reasoning quality. We begin with high-recall lexical retrieval, using controlled query expansion to address vocabulary mismatch (Miller, 1995). RRF then stabilizes the combination of diverse query variants (Cormack et al., 2009). A precision-oriented cross-encoder re-ranks before any generation, consistent with standard multi-stage practice on MS MARCO/TREC-style setups (Nogueira and Cho, 2019; Robertson and Zaragoza, 2009). Finally, the agentic draft-and-critique loop imposes separation of concerns and structured self-feedback, reducing "LLM-grades-

[1] See the appendix for aggregated context

its-own-paper" failure modes and improving report quality (Wataoka et al., 2024; Li et al., 2024; Madaan et al., 2023; Glass et al., 2022).

4 Related Work

Research on automated news trustworthiness assessment has largely centered on fact-checking, retrieval-based verification, and retrieval-augmented generation. Early efforts, exemplified by the FEVER benchmark (Thorne et al., 2018), evaluated systems that verified individual claims against curated textual sources such as Wikipedia. In this context, a claim is a short, standalone factual statement that can be checked against textual evidence.. These systems performed well on isolated claims but struggled with full news articles, which often require connecting information from multiple sources and assessing the reliability or bias of those sources. In addition, traditional fact-checking systems generally produce true or false labels, which provide limited help for readers trying to understand the broader context of available information.

The TREC 2024 Lateral Reading Track (Zhang et al., 2024) addressed some of these limitations by introducing verification strategies modeled after how people evaluate information, where readers look at outside sources to check a page's reliability, a major contrast to the previous work which focused on information within the source itself. The track involved two main tasks: generating investigative questions and retrieving evidence to answer them. Its evaluation prioritized reasoning quality over retrieval accuracy, encouraging systems to focus on relevance, completeness, and inclusion of different viewpoints. The results of this track showed that the combination of question generation and retrieval can more closely match how people verify information, leading to more thorough evaluations. According to the track conclusion, stronger systems went beyond prompting LLMs alone by incorporating strategies such as query expansion for document retrieval and the use of agents in workflows using online resources to gather more context for question generation, all of which improved overall performance (Zhang et al., 2024).

Retrieval-augmented generation (RAG) frameworks (Lewis et al., 2020) build on these ideas by allowing language models to generate text based on retrieved evidence while maintaining explicit source attribution. This use of source-attribution helps reduce hallucination and keeps connections between evidence and claims clear to the reader. The TREC 2024 RAG Track further explored this by evaluating how retrieval and generation components could be integrated. For instance, the Ragnarök framework (Pradeep et al., 2024) was selected to provide the official baselines for the track, serving as a standardized foundation for evaluating retrieval-augmented systems. It was designed to support MS MARCO V2.1, a large and diverse document collection, and to support research on complex questions that require breaking problems into parts and providing detailed and well-reasoned answers. Many participating teams explored multi-stage pipelines that combined dense retrievers, rerankers, and instruction-tuned language models to improve factual grounding and coherence (TREC Retrieval Group, 2024). Reported results suggest that separating retrieval, summarization, and attribution stages offered practical advantages for precision and interpretability over single-stage, end-to-end approaches.

The DRAGUN track builds on these ideas. While the RAG 2024 task focused on ensuring answers were factually correct and complete at the nugget level in open-domain questions, DRAGUN applies the same principles to understanding news articles. Both tracks aim to validate that generated outputs are backed by verifiable evidence, but DRAGUN goes further by requiring clear, citation-supported reports that help readers make sense of conflicting or incomplete information. In this way, DRAGUN continues the path started by RAG 2024 by emphasizing systems that combine evidence from multiple sources and presenting it in a way that is easy to follow, similar to human fact-checking.

5 Conclusion

CITADEL was designed to test whether a relatively simple, classical IR backbone combined with an agentic draft–critique loop could meet the DRAGUN goal of producing grounded, citation-rich news reports. Our pipeline uses BM25 with structured synonym expansion, RRF fusion, and cross-encoder reranking to surface a compact but diverse set of candidate passages, and then delegates to separate generator and evaluator agents to iteratively refine a report against an explicit rubric. Within the DRAGUN evaluation setup—where NIST assessors generate investigative questions and score

systems based on how many questions are fully or partially answered—this design performed well: based on organizer-provided per-topic statistics, our average topic score was approximately twice the median across participants. Because the scoring function is directly tied to coverage of assessor questions, this suggests that the combination of high-recall retrieval with a multi-round, rubric-driven loop improves the breadth of issues a report addresses.

Preliminary error analysis points to systematic strengths and weaknesses. On topics where most assessor questions focused on the substantive content of the article and where relevant evidence was well represented in the MS MARCO corpus, our system tended to perform strongly. In these cases, the IR pipeline typically surfaced at least one high-quality passage for each major sub-issue, and the generator–evaluator loop was able to turn that evidence into dense, well-cited coverage. In contrast, we likely underperformed on topics where many questions targeted author- or source-level credibility (e.g., the outlet's track record, funding, or political alignment). Our prompts and tools explicitly prioritized the content of the claims in the article rather than the provenance of the author or outlet, and the retrieval layer was not tuned to search for biographical or institutional profiles. Some questions may also have been missed because the retrieval stage failed to surface the right kind of evidence or because the LLM overlooked less salient aspects when summarizing long contexts.

One design choice that may have given CITADEL an advantage is how we couple generation and question-asking. DRAGUN includes an optional shared task in which systems submit a fixed set of "best" questions for fact-checking each document; a natural overall pipeline is then questions $\rightarrow$ answers $\rightarrow$ report. By contrast, our system starts from a draft report and uses the evaluator's rubric and free-text notes to generate follow-up questions on the fly based on observed gaps. We also do not cap the number of questions at 10: in an eight-round run, the evaluator typically proposes on the order of 20–40 questions, which are distributed to IR agents and folded back into an expanding context document. This summary-first, question-on-demand pattern appears to help coverage, because questions are directly conditioned on what is currently missing from the report rather than on a one-shot guess at what will matter.

At the same time, our approach has clear limitations. On the retrieval side, we rely on a purely lexical Lucene index with WordNet-based expansion and a single cross-encoder stage; compute and storage constraints prevented us from building a dense or fully hybrid index over the entire corpus. This keeps the system simple and reproducible but leaves potential gains from modern dense retrievers on the table, especially for semantically subtle or paraphrased evidence. Azure and Cohere quota constraints also forced us to prune aggressively: only a small subset of candidate passages and metadata can be passed into the agentic layer per round, and long documents must be compressed into summaries. These bottlenecks make the final report sensitive to early retrieval and summarization choices.

The agentic layer inherits its own constraints. We use GPT-4.1 for both report generation and evaluation, separated into distinct roles but sharing the same underlying model family. This role separation reduces some "LLM grades its own paper" failure modes, but it does not eliminate shared blind spots or calibration errors. Our evaluator operates under a fixed, hand-designed rubric and a small number of rounds, so it may systematically underweight certain dimensions (such as provenance or contradiction across sources) that DRAGUN assessors care about. Practical factors such as limited development time and the need to stay within rate limits also restricted the extent of prompt tuning, ablation, and robustness testing we could perform.

Despite these constraints, our DRAGUN results suggest that (1) classical, transparent IR methods remain highly competitive when paired with a strong reranker, and (2) test-time, rubric-driven critique with dynamically generated questions is a promising pattern for retrieval-augmented news understanding. Future work could integrate dense or hybrid retrievers, add dedicated tools for author and outlet profiling, diversify evaluators across model families, and explore learned or adaptive stopping criteria for the draft–critique loop. More broadly, DRAGUN provides a valuable testbed for studying how systems like CITADEL can help readers interrogate news at scale, not by declaring verdicts, but by organizing evidence, highlighting uncertainties, and making the link between claims and citations as explicit as possible.

6 Post Hoc Evaluation

6.1 Task 2 Evaluation Protocol

DRAGUN Task 2 is evaluated via question-based assessment rather than traditional retrieval metrics. For each topic (a real news article), NIST assessors write investigative questions that may target both the article's substantive claims and broader credibility context (e.g., missing background, alternative viewpoints, or source reliability). A system's 250-word report is then judged according to how well it addresses these assessor questions using evidence from the MS MARCO V2.1 segmented corpus, with explicit passage-level citations (Zhang et al., 2025). In organizer-provided score breakdowns, performance is reported as per-run averages with uncertainty estimates (95% confidence intervals), and scores are decomposed into *supportive* and *contradictory* components reflecting whether report content aligns with (or conflicts with) assessor judgments.

This evaluation setup strongly rewards *coverage*: systems that surface and address a broad set of highly relevant investigative angles—while keeping claims grounded in cited evidence—tend to score better than systems that only answer a narrow subset of questions exceptionally well.

6.2 Main Results

Figure 4 summarizes organizer-reported per-run average scores (with 95% confidence intervals). Our `[SCIAI] SCIAI_03_. . .` runs fall among the top-performing groups for Task 2 by supportive score while maintaining low contradictory scores. Qualitatively, the gap between our baseline configuration and the iterative `SCIAI_03` loop variants suggests that much of our gain is attributable to iterative refinement (rubric-based critique + targeted follow-up questions), rather than to changes in first-stage retrieval.

6.3 Why Iterative Questioning Helps Task 2

A central design choice in CITADEL is that we treat "question generation" as an *internal control signal* rather than as a required precursor to report writing. DRAGUN includes an optional Task 1 in which systems submit up to 10 investigative questions per topic. A natural end-to-end design is therefore:

topic → generate questions (Task 1)
→ retrieve/answer
→ write report (Task 2).

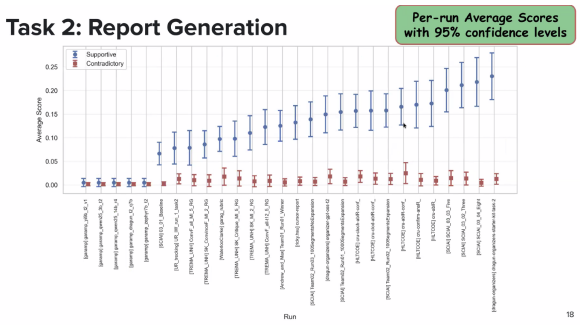

Figure 4: Organizer-provided DRAGUN Task 2 per-run average scores with 95% confidence intervals, separated into supportive and contradictory components. Our submissions correspond to the `[SCIAI] SCIAI_03_. . .` runs.

We intentionally did not optimize for Task 1 submission quality. Instead, our evaluator produces follow-up questions dynamically during report refinement, conditioned on observed gaps in coverage and weakly supported statements in the current draft. This has two practical consequences for Task 2:

- **Higher question volume.** The evaluator typically proposes many more than 10 questions over multiple rounds (often dozens), each tied to a specific missing facet of the report. This increases the chance that at least one query formulation will surface evidence relevant to an assessor question.

- **Adaptivity to retrievability.** When a proposed investigative angle is difficult to answer given corpus coverage, the evaluator can pivot to nearby, more retrievable formulations (e.g., alternative names, institutions, timelines, or counterclaims). This reduces the likelihood of spending the entire budget on a small set of unanswerable questions.

This strategy is compatible with the organizer starter kit's strong overall performance while still producing Task 1 outputs: in that pipeline, evidence collection is iterated first, and questions are generated *after* sufficient information has been gathered (i.e., retrospectively rather than prospectively). In other words, strong Task 2 performance does not require treating Task 1 as a strict upstream dependency; both tasks can be driven by a shared iterative information-gathering stage.

6.4 Post Hoc Analysis Constraints

Our ability to interpret per-topic outcomes is limited by the information released with the evaluation. In particular, we did not receive the assessor question sets used to grade reports; we received only the resulting scores. As a consequence, our analysis of *why* a topic scored well or poorly is necessarily post hoc and partially speculative, based on the generated reports, retrieved evidence, and our own system traces rather than on direct alignment to the assessor questions.

In addition, at the time of writing, most participating teams' notebook papers and full system descriptions have not yet been published. This prevents us from attributing our relative performance to specific methodological differences (e.g., dense retrieval, alternative rerankers, or different agent decompositions) beyond what is described in organizer materials and public baselines.

6.5 Error Analysis and Threats to Validity

We inspected topic-level scores and found that our best-performing topics were `msmarco_v2.1_doc_04_420132660` and `msmarco_v2.1_doc_35_441326441`, while our worst-performing topics were `msmarco_v2.1_doc_25_502424913` and `msmarco_v2.1_doc_15_116067546`. Without assessor questions, we cannot conclusively identify the dimensions driving these extremes; however, several recurring patterns emerged from manual review of our system outputs.

First, our strongest topics tend to be those where assessor questions plausibly focus on the *substantive content* of the article and where relevant corroborating material is readily surfaced in the MS MARCO corpus. In these cases, high-recall retrieval plus multi-round critique is effective: the IR layer typically surfaces at least one strong passage per sub-issue, and the generator/evaluator loop produces dense, well-cited coverage.

Second, we expect weaker performance on topics where assessor questions emphasize *provenance* (e.g., outlet reputation, author background, funding, institutional incentives, or long-run credibility signals). Our prompts and tools primarily target claim-level content and contextual background rather than explicit profiling of sources or authors, and our retrieval layer was not tuned specifically for biographical or institutional "who is this outlet/author" queries. This is a deliberate prioritization of content validity over source/author validity; if a topic is graded heavily on provenance, our approach is likely disadvantaged.

Third, we observed that our reports can become susceptible to the *topic article's framing* when the article is persuasive but contains partially inaccurate or weakly supported claims. A concrete example is `msmarco_v2.1_doc_15_116067546`, which frames quantum computing as an immediate threat to Bitcoin. In our system trace for this topic, we retrieved passages indicating that quantum computers do not currently pose a near-term practical risk to Bitcoin (e.g., due to hardware scale requirements, error correction constraints, and the feasibility of protocol migration). Nevertheless, the generated report tended to mirror the article's urgency framing and emphasis, treating the threat as more immediate than warranted by the retrieved evidence. This failure mode is exacerbated when counterevidence is less salient than the topic narrative, or when retrieval yields mitigating details that are technically correct but harder to incorporate into a concise report under strict length and citation constraints.

Finally, we found that the LLM sometimes cites the topic document itself in the report. While this can be useful for attributing what the article claims, it is not necessarily useful evidence for assessing *whether* the claims are true. Over-reliance on citing the topic document may reduce the effective diversity of corroborating sources and can weaken the report's ability to demonstrate independent verification.

Beyond these topic-level patterns, several practical constraints limit interpretability and suggest straightforward avenues for improvement:

- **Retrieval limitations.** We rely on a lexical Lucene index with WordNet-based expansion and a single cross-encoder re-ranking stage; we did not build a dense/hybrid index over the full collection due to compute and storage constraints.

- **Budget and truncation effects.** Azure/Cohere quotas require aggressive pruning (e.g., limiting how much retrieved text and metadata can be passed downstream per round), making outcomes sensitive to early retrieval and summarization choices.

- **Shared model family for generation and critique.** Although we separate generator and

evaluator roles, both agents use the same underlying model family, which can preserve shared blind spots and calibration errors (a partial instance of "LLM grades its own paper" failure modes).

Despite these limitations, the overall results support the main claim of the paper: pairing a transparent, recall-oriented IR stack with a rubric-driven draft–evaluate loop yields strong Task 2 performance in the DRAGUN setting, especially on the core objective of producing broad, citation-grounded coverage of assessor investigative questions.

7 Acknowledgments

We gratefully acknowledge the support of the Center for Undergraduate Research and Creative Activity (CURCA) at Siena University, which provided funding for this research. Their assistance made this project possible.

References

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, Boston, MA, USA. Association for Computing Machinery (ACM).

Michael R. Glass, Shiyue Zhang, Denis Savenkov, Alfio Gliozzo, J. Scott McCarley, J. William Murdock, and Achille Fokoue. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2662–2678, Seattle, United States. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Vladimir Karpukhin, Naman Goyal, Dhruv Batra, Devi Parikh, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33.

Tianyi Li, Yujia Liu, Zhiyuan Zhang, Peng Xu, Rui Wang, Wei Li, Weizhe Chen, and Maosong Sun. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Jimmy Lin, Xueguang Ma, and Rodrigo Nogueira. 2024. Pyserini lucene index for ms marco v2.1 (segmented full). `https://rgw.cs.uwaterloo.ca/pyserini/indexes/lucene/lucene-inverted.msmarco-v2.1-doc-segmented-full.20240418.4f9675.tar.gz`. Version 20240418, retrieved from the Pyserini GitHub/Waterloo RGW index repository.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Ronak Pradeep, Nandan Thakur, Nils Reimers, Charles LA Clarke, and Jimmy Lin. 2024. Ragnarök: A reusable rag framework and baselines for the trec 2024 retrieval-augmented generation track. *arXiv preprint arXiv:2406.16828*.

Princeton University. 2010. Wordnet: A lexical database for english. `https://wordnet.princeton.edu/download`. Version 3.0, accessed 2025-10-28.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

TREC Retrieval Group. 2024. Trec 2024 retrieval-augmented generation (rag) retrieval task appendix. Accessed: 2025-10-22.

Yusuke Wataoka, Ryo Takahashi, and Ryosuke Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. 2024. Overview of the trec 2024 lateral reading track. In *Proceedings of the 33rd Text REtrieval Conference (TREC 2024)*, Gaithersburg, MD, USA. National Institute of Standards and Technology (NIST).

Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. 2025. Trec 2025 detection, retrieval, and augmented generation for understanding news (dragun) track organizers. https://trec-dragun.github.io/#organizers. Accessed: 2025-06-18.

A Full Two-Round Trace

Topic 1. (as provided by DRAGUN)

Generator–0 JSON (no IR).

Listing 1: Generator–0 CoT

```
<cot>
My plan is to produce a concise and information-
    dense report summarizing the origins and
    cultural significance of Hawaiian pizza,
    based on the provided article. I will
    prioritize the most critical points: the
    inventor (Sam Panopoulos), the location and
    date of invention (Ontario, Canada, 1962),
    the ingredients and initial reception, and
    the broader context of pizza's arrival in
    North America. I will also note the
    controversy and divisiveness of Hawaiian
    pizza, as well as its fusion food roots.
    Each text block will be self-contained,
    neutral, and factually accurate, referencing
     the article as needed. I will ensure the
    total word count across all text blocks does
     not exceed 250, and the output will be
    valid JSON as required.
</cot>
```

Listing 2: Generator–0 Note

```
<note>
This is the first round, so all information is
    sourced from the provided article. There are
     no external IR citations yet; all citations
     will be left empty. Coverage focuses on the
     invention, cultural context, and
    controversy of Hawaiian pizza. No questions
    were left unanswered from the input.
</note>
```

Listing 3: Generator–0 Report

```
<report>
{
   "responses": [
      {
         "text": "Hawaiian pizza was invented
             in 1962 by Sam Panopoulos, a
             Greek-Canadian restaurateur in
             Ontario, Canada. Despite its name
             , the pizza has no direct
             connection to Hawai'i and was
             created in a diner serving
             typical mid-century food. The
             dish combines canned pineapple
             and ham atop a cheese pizza, and
             its creation was influenced by
             the popularity of canned
             pineapple and tiki culture in
             North America during the 1960s.",
         "citations": []
      },
      {
         "text": "Panopoulos began
             experimenting with pizza after
             observing its preparation in
             Windsor, Canada, and Detroit, USA.
              At the time, pizza was a
             relatively new and unfamiliar
             dish in Canada, with limited
             topping options and primitive
             cooking methods. The addition of
             pineapple was inspired by the
             sweet and savory flavors found in
              Chinese cuisine, which
             Panopoulos already served at his
             restaurant.",
         "citations": []
      },
      {
         "text": "Since its inception,
             Hawaiian pizza has become a
             divisive and controversial item
             in the pizza world, often
             polarizing opinions among
             consumers. While the classic
             combination of ham and pineapple
             was initially accidental,
             Panopoulos named the creation '
             Hawaiian pizza' from the start,
             and it quickly gained popularity
             despite its unconventional
             ingredients.",
         "citations": []
      }
   ]
}
</report>
```

Evaluator–0 JSON (rubric + questions).

Listing 4: Evaluator–0 CoT

```
<cot>
I will grade the report based on coverage of the
     invention, cultural context, and
    controversy of Hawaiian pizza, as described
    in the topic document. I will check for
    information density, completeness, and
    prioritization of key facts (e.g., inventor,
     date, location, influences, controversy).
    Citation quality will be marked down since
    no citations are provided, even though the
    generator notes this is due to lack of IR
    context. Style will be evaluated for clarity
     and conciseness. I will probe for gaps such
     as more detail on Panopoulos's background,
    the evolution of the pizza's recipe, and its
     reception over time. I will ensure my
    answer is valid JSON in all required fields.
</cot>
```

Listing 5: Evaluator–0 Note

```
<note>
Please add more specific details about
    Panopoulos's background and the Satellite
    restaurant, as these are mentioned in the
    source and add depth. Consider including
    direct quotes or more granular facts (e.g.,
    the process of pizza box improvisation, the
    initial toppings available) to increase
```

```
    information density. When IR context is
    available, cite specific segments to
    strengthen reliability. The word count for
    all text blocks is under 250.
</note>
```

Listing 6: Evaluator–0 Questions

```
<ir>
{
  "questions": [
    {
      "question": "What was the initial public
          and media reaction to Hawaiian pizza
          after its invention?",
      "context": "The article mentions
          controversy and divisiveness but does
          not detail early reactions."
    },
    {
      "question": "How did the Satellite
          restaurant market and serve Hawaiian
          pizza in its early years?",
      "context": "Panopoulos began advertising
          his new pizza topping, but details are
           sparse."
    },
    {
      "question": "Are there any records or
          interviews with customers or staff
          from the Satellite restaurant
          regarding the invention?",
      "context": "The article relies on
          Panopoulos's recollections; external
          perspectives are missing."
    },
    {
      "question": "How did the recipe for
          Hawaiian pizza evolve after its
          initial creation?",
      "context": "It is noted that the ham and
          pineapple combination was accidental,
          but further evolution is not described
          ."
    }
  ]
}
</ir>
```

Listing 7: Evaluator–0 Rubric Scores

```
<eval>
{
  "coverage": 4,
  "accuracy": 5,
  "citation_quality": 2,
  "style": 4,
  "prioritization": 4,
  "completeness": 3,
  "information_density": 3,
  "comments": "The report accurately summarizes
      the invention and cultural context of
      Hawaiian pizza, prioritizing key facts.
      However, it lacks granular details and
      direct quotes that would increase
      information density and completeness.
      Citation quality is low due to absence of
      references, though this is understandable
      for the first round. Coverage is strong
      but could be improved with more specifics
      about Panopoulos and the Satellite
      restaurant. Style is clear and concise."
}
</eval>
```

IR Fan-out JSON (per-question answers + citations).

Listing 8: IR-Agent Search (CoT)

```
<cot>
For each question, I will generate up to 4 BM25-
    optimized keyword queries targeting specific
     aspects, and a master query for semantic
    reranking. The queries will focus on public/
    media reaction and restaurant marketing/
    serving methods. I reaffirm that my answer
    will be valid JSON.
</cot>
```

Listing 9: IR-Agent Search (Answer payload)

```
<answer>
{
    "searches": [
        {
            "queries": [
                "initial public reaction to
                    Hawaiian pizza invention",
                "media coverage of Hawaiian pizza
                    after it was created",
                "controversy surrounding Hawaiian
                    pizza in its early years",
                "reception of pineapple pizza by
                    public and press"
            ],
            "master_query": "What was the initial
                 public and media reaction to
                Hawaiian pizza after its
                invention, including any
                controversy or divisiveness?"
        },
        {
            "queries": [
                "how Satellite restaurant marketed
                     Hawaiian pizza",
                "early advertising strategies for
                    Hawaiian pizza by Satellite
                    restaurant",
                "methods used to serve Hawaiian
                    pizza at Satellite restaurant
                    ",
                "Panopoulos promotion of pineapple
                     pizza in the restaurant's
                    early years"
            ],
            "master_query": "How did the
                Satellite restaurant market and
                serve Hawaiian pizza in its early
                 years, including advertising and
                 serving methods?"
        }
    ]
}
</answer>
```

Listing 10: IR-Agent Select (CoT)

```
<cot>
I reviewed the search metadata for relevance to
    the questions about initial public/media
    reaction to Hawaiian pizza and how the
    Satellite restaurant marketed and served it.
     I selected segments that likely contain
    historical context, controversy, marketing,
    and inventor details. I reaffirm that my
    answer will be valid JSON.
</cot>
```

Listing 11: IR-Agent Select (Answer payload)

```
<answer>
{
  "selections": [
    "msmarco_v2.1_doc_28_903296016#2_1908541802",
    "msmarco_v2.1_doc_30_523003770#0_1183664316",
    "msmarco_v2.1_doc_57_619272585#9_900954951",
    "msmarco_v2.1_doc_51_759827373#0_1589150899",
    "msmarco_v2.1_doc_16_565495028#0_805615245",
    "msmarco_v2.1_doc_17_1448273400#0_1598351114"
  ]
}
</answer>
```

Listing 12: IR-Agent Update (CoT)

```
<cot>
Using the provided segments, I can confirm that
    the initial public and media reaction to
    Hawaiian pizza was divisive and
    controversial, but detailed early reactions
    are not extensively documented. The pizza
    was invented by Sam Panopoulos in 1962 at
    the Satellite Restaurant in Ontario, Canada,
     where he and his brothers experimented with
     pineapple on pizza to attract customers.
    The pizza was named after the brand of
    canned pineapple used. Early popularity was
    limited but grew locally and then globally.
    The Satellite restaurant marketed it by
    experimenting with new flavors, but specific
     marketing strategies are not detailed. I
    reaffirm that my answer will be valid JSON.
</cot>
```

Listing 13: IR-Agent Update (Answers + rounds)

```
<answer>
{
  "questions": [
    {
      "question": "What was the initial public
          and media reaction to Hawaiian pizza
          after its invention?",
      "doc_context": "The article mentions
          controversy and divisiveness but does
          not detail early reactions.",
      "answer": {
        "text": "The initial public and media
            reaction to Hawaiian pizza was
            divisive and controversial, with the
            pizza inspiring jokes, memes, and
            debates. However, detailed early
            reactions are not extensively
            documented. The pizza was invented in
             1962 by Sam Panopoulos, who defended
             his creation throughout his life.
            While the addition of pineapple was
            not initially very popular, it
            eventually became a staple offering.
            The controversy mainly centers on the
             inclusion of pineapple as a topping
            .",
        "citations": [
          {
            "summary": "Describes the
                divisiveness and controversy
                surrounding Hawaiian pizza and
                its inventor Sam Panopoulos.",
            "citation": "msmarco_v2.1
                _doc_28_903296016#2_1908541802"
          },
          {
            "summary": "Details the invention of
                Hawaiian pizza by Sam Panopoulos
                in 1962 and the naming after the
                canned pineapple brand.",
            "citation": "msmarco_v2.1
                _doc_16_565495028#0_805615245"
          },
          {
            "summary": "Notes the controversy and
                 polarizing debate mainly focused
                 on pineapple as a topping.",
            "citation": "msmarco_v2.1
                _doc_17_1448273400#0_1598351114"
          }
        ]
      },
      "finished": true
    },
    {
      "question": "How did the Satellite
          restaurant market and serve Hawaiian
          pizza in its early years?",
      "doc_context": "Panopoulos began
          advertising his new pizza topping, but
           details are sparse.",
      "answer": {
        "text": "The Satellite restaurant, where
            Hawaiian pizza was invented by Sam
            Panopoulos in 1962, marketed the
            pizza by experimenting with new
            flavor combinations to attract
            customers. Panopoulos and his
            brothers added pineapple and ham to a
             standard cheese pizza as a fun
            experiment. The pizza was named after
             the canned pineapple brand used.
            Specific marketing strategies or
            serving methods beyond this
            experimentation are not detailed in
            the available information.",
        "citations": [
          {
            "summary": "Describes Panopoulos'
                experimentation with pineapple
                and ham at the Satellite
                restaurant to attract customers
                .",
            "citation": "msmarco_v2.1
                _doc_30_523003770#0_1183664316"
          },
          {
            "summary": "Notes the invention of
                Hawaiian pizza at the Satellite
```

```
                Diner and its naming after the
                canned pineapple brand.",
            "citation": "msmarco_v2.1
                _doc_57_619272585#9_900954951"
          },
          {
            "summary": "Mentions Panopoulos'
                experience and experimentation at
                 the Satellite restaurant in
                Ontario, Canada.",
            "citation": "msmarco_v2.1
                _doc_16_565495028#0_805615245"
          }
        ]
      },
      "finished": true
    }
  ],
  "rounds": [
    {
      "summary": "Initial round focused on
          queries about public/media reaction
          and Satellite restaurant marketing.
          Found relevant segments confirming
          invention details, divisiveness, and
          experimentation but no detailed early
          marketing strategies or media
          reactions.",
      "seen_ids": [
        "msmarco_v2.1_doc_28_903296016#2
            _1908541802",
        "msmarco_v2.1_doc_30_523003770#0
            _1183664316",
        "msmarco_v2.1_doc_57_619272585#9
            _900954951",
        "msmarco_v2.1_doc_16_565495028#0
            _805615245",
        "msmarco_v2.1_doc_17_1448273400#0
            _1598351114"
      ]
    }
  ]
}
</answer>
```


Full Context Doc (per-question stitched answers).

Listing 14: Context Q1


```
{
"question": "Are there any records or interviews
     with customers or staff from the Satellite
    restaurant regarding the invention?",
"doc_context": "The article relies on Panopoulos
    's recollections; external perspectives are
    missing.", "answer": {
   "text": "There are no known records or
       interviews with customers or staff from
       the Satellite restaurant regarding the
       invention of Hawaiian pizza. The
       information primarily relies on Sam
       Panopoulos's own recollections and
       statements.",
   "citations": [
      {
      "summary": "Panopoulos created Hawaiian
          pizza at the Satellite Restaurant in
          1962, but external perspectives are
          missing.", "citation": "msmarco_v2.1
          _doc_04_420132660#5_965977302"
      },
      {
      "summary": "Panopoulos described the
          invention and experimentation with
          pineapple on pizza, with no mention
          of other interviews.",
      "citation": "msmarco_v2.1
          _doc_30_523003770#0_1183664316"
      }
   ]},
"finished": true
}
```


Listing 15: Context Q2


```
{
"question": "How did the recipe for Hawaiian
    pizza evolve after its initial creation?",
"doc_context": "It is noted that the ham and
    pineapple combination was accidental, but
    further evolution is not described.",
"answer":
   {
   "text": "The recipe for Hawaiian pizza was
       initially an experiment by Sam
       Panopoulos, inspired by mixing sweet and
        savory flavors common in Chinese dishes.
        It started with ham and pineapple,
       named after the brand of canned
       pineapple used. While initially not very
        popular, it gradually became a staple
       offering worldwide. However, detailed
       information on how the recipe evolved
       after its creation is not provided.",
   "citations": [
      {
      "summary": "Panopoulos experimented with
          pineapple and ham inspired by sweet
          and savory flavor combinations,
          naming it Hawaiian after the
          pineapple brand.", "citation": "
          msmarco_v2.1_doc_16_565495028#0
          _805615245"
      },
      {
      "summary": "The addition of pineapple and
           ham was initially unpopular but
          became popular locally and globally
          over time.", "citation": "msmarco_v2
          .1_doc_17_1448273400#0_1598351114"}
   ]},
"finished": true
}
```


Listing 16: Context Q3


```
{
"question": "What was the initial public and
    media reaction to Hawaiian pizza after its
    invention?", "doc_context": "The article
    mentions controversy and divisiveness but
    does not detail early reactions.", "answer":
   {
   "text": "The initial public and media
       reaction to Hawaiian pizza was divisive
       and controversial, with the pizza
       inspiring jokes, memes, and debates.
       However, detailed early reactions are
```

```
        not extensively documented. The pizza
        was invented in 1962 by Sam Panopoulos,
        who defended his creation throughout his
         life. While the addition of pineapple
        was not initially very popular, it
        eventually became a staple offering. The
         controversy mainly centers on the
        inclusion of pineapple as a topping.", "
        citations": [
       {
       "summary": "Describes the divisiveness
           and controversy surrounding Hawaiian
           pizza and its inventor Sam Panopoulos
           .",
       "citation": "msmarco_v2.1
           _doc_28_903296016#2_1908541802"
       },
       {
       "summary": "Details the invention of
           Hawaiian pizza by Sam Panopoulos in
           1962 and the naming after the canned
           pineapple brand.", "citation": "
           msmarco_v2.1_doc_16_565495028#0
           _805615245"
       },
       {
       "summary": "Notes the controversy and
           polarizing debate mainly focused on
           pineapple as a topping.", "citation":
            "msmarco_v2.1_doc_17_1448273400#0
           _1598351114"}
   ]},
"finished": true
}
```

Listing 17: Context Q4

```
{
"question": "How did the Satellite restaurant
    market and serve Hawaiian pizza in its early
     years?", "doc_context": "Panopoulos began
    advertising his new pizza topping, but
    details are sparse.",
"answer":
   {
   "text": "The Satellite restaurant, where
       Hawaiian pizza was invented by Sam
       Panopoulos in 1962, marketed the pizza
       by experimenting with new flavor
       combinations to attract customers.
       Panopoulos and his brothers added
       pineapple and ham to a standard cheese
       pizza as a fun experiment. The pizza was
        named after the canned pineapple brand
       used. Specific marketing strategies or
       serving methods beyond this
       experimentation are not detailed in the
       available information.",
   "citations": [
       {
       "summary": "Describes Panopoulos'
           experimentation with pineapple and
           ham at the Satellite restaurant to
           attract customers.",
       "citation": "msmarco_v2.1
           _doc_30_523003770#0_1183664316"
       },
       {
       "summary": "Notes the invention of
           Hawaiian pizza at the Satellite Diner
            and its naming after the canned
           pineapple brand.", "citation": "
           msmarco_v2.1_doc_57_619272585#9
           _900954951"
       },
       {
       "summary": "Mentions Panopoulos'
           experience and experimentation at the
            Satellite restaurant in Ontario,
           Canada.", "citation": "msmarco_v2.1
           _doc_16_565495028#0_805615245"
       }
   ]},
"finished": true
}
```

Generator–1 JSON (revised with citations).

Listing 18: Generator–1 CoT

```
<cot>
To create a concise, information-dense report, I
     will first summarize the invention and
    origins of Hawaiian pizza, emphasizing Sam
    Panopoulos's background and the context of
    the Satellite restaurant. Next, I will
    address the lack of external records or
    interviews about the invention, relying on
    Panopoulos's own recollections. I will then
    detail the evolution of the recipe, noting
    the initial experiment and subsequent
    popularity, while acknowledging gaps in
    documentation. Finally, I will discuss the
    initial public and media reaction,
    highlighting controversy and divisiveness,
    and describe how the Satellite restaurant
    marketed and served the pizza in its early
    years. All claims will be attributed to the
    relevant segment_ids from the IR context. I
    will ensure the JSON output is valid and the
     total word count for all text blocks is
    under 250 words.
</cot>
```

Listing 19: Generator–1 Note

```
<note>
Evaluator: This round incorporates IR context,
    with citations for each factual claim.
    Details about Panopoulos's background, the
    Satellite restaurant, recipe evolution, and
    public reaction are included. Where
    information is sparse (e.g., marketing
    strategies, external interviews), gaps are
    acknowledged and attributed. No questions
    were left unanswered; all available IR
    context has been used. Word count is within
    the limit.
</note>
```

Listing 20: Generator–1 Report

```
<report>
{
   "responses": [
      {
          "text": "Hawaiian pizza was invented
              in 1962 by Sam Panopoulos, a
              Greek-Canadian restaurateur, at
```

```
                the Satellite restaurant in
                Ontario, Canada. The dish
                combined canned pineapple and ham
                 atop a cheese pizza, inspired by
                 Panopoulos's interest in mixing
                sweet and savory flavors, a
                concept familiar from Chinese
                cuisine served at his diner.",
            "citations": ["msmarco_v2.1
                _doc_04_420132660#5_965977302", "
                msmarco_v2.1_doc_16_565495028#0
                _805615245", "msmarco_v2.1
                _doc_30_523003770#0_1183664316"]
        },
        {
            "text": "There are no known records
                or interviews with customers or
                staff from the Satellite
                restaurant regarding the
                invention of Hawaiian pizza; most
                 information relies on Panopoulos
                's own recollections and
                statements.",
            "citations": ["msmarco_v2.1
                _doc_04_420132660#5_965977302", "
                msmarco_v2.1_doc_30_523003770#0
                _1183664316"]
        },
        {
            "text": "The recipe for Hawaiian
                pizza began as an experiment with
                 pineapple and ham, named after
                the canned pineapple brand used.
                While initially unpopular, it
                gradually gained local and global
                 popularity, though detailed
                documentation of its evolution is
                 lacking.",
            "citations": ["msmarco_v2.1
                _doc_16_565495028#0_805615245", "
                msmarco_v2.1_doc_17_1448273400#0
                _1598351114"]
        },
        {
            "text": "The initial public and media
                 reaction to Hawaiian pizza was
                divisive and controversial,
                sparking debates and jokes,
                mainly focused on the inclusion
                of pineapple as a topping.
                Panopoulos defended his creation
                throughout his life.",
            "citations": ["msmarco_v2.1
                _doc_28_903296016#2_1908541802",
                "msmarco_v2.1_doc_16_565495028#0
                _805615245", "msmarco_v2.1
                _doc_17_1448273400#0_1598351114"]
        }
    ]
}
</report>
```

Evaluator–1 JSON (final rubric + follow-ups).

Listing 21: Evaluator–1 CoT

```
<cot>
I will evaluate the report for coverage,
    accuracy, citation quality, style,
    prioritization, completeness, and
    information density. I will check if all
    major aspects from the topic and IR context
    are addressed, especially details about
    Panopoulos's background, the Satellite
    restaurant, the invention process, and the
    cultural context/controversy. I will verify
    that citations are relevant and specific,
    and that the report avoids redundancy.
    Information density will be graded based on
    how many unique, granular facts are included
    . I will ensure my output is valid JSON and
    within the word count limit. Scores will
    reflect the report's strengths in
    summarizing available evidence, but will be
    lower where details are sparse or citations
    are not tightly linked to claims. I will
    probe for missing details about the
    Satellite restaurant's operations, marketing
    , and the broader context of pizza in Canada
     at the time.
</cot>
```

Listing 22: Evaluator–1 Note

```
<note>
Please increase information density by including
     more granular details from the source, such
     as Panopoulos's improvisation with pizza
    boxes, the types of toppings available in
    early Canadian pizza, and the context of
    pizza's novelty in Ontario. Where possible,
    use direct quotes or specific anecdotes to
    strengthen the narrative and citation
    reliability. Consider clarifying the
    timeline and process of how Hawaiian pizza
    became popular locally and globally, as this
     is currently vague. The total word count
    for all text blocks is under 250.
</note>
```

Listing 23: Evaluator–1 New Questions

```
<ir>
{
  "questions": [
    {
      "question": "What specific anecdotes or
          direct quotes from Panopoulos or
          others exist regarding the invention
          and early reception of Hawaiian pizza
          ?",
      "context": "The report relies on Panopoulos
          's recollections but does not include
          direct quotes or anecdotes, which
          could add depth and reliability."
    },
    {
      "question": "How did the Satellite
          restaurant adapt its menu and
          operations to accommodate the
          introduction of pizza, and what was
          the customer response?",
      "context": "Details about the Satellite
          restaurant's adaptation and customer
          response are missing; only general
          statements about experimentation are
          provided."
    },
    {
```

```
    "question": "What was the broader context
        of pizza's introduction to Ontario and
         Canada in the 1950s and 1960s, and
        how did it differ from U.S. pizza
        culture?",
    "context": "The article mentions pizza's
        novelty in Ontario and differences
        from Detroit, but lacks specifics
        about the broader Canadian context."
  }
 ]
}
</ir>
```

Listing 24: Evaluator–1 Rubric Scores

```
<eval>
{
 "coverage": 4,
 "accuracy": 5,
 "citation_quality": 4,
 "style": 4,
 "prioritization": 4,
 "completeness": 3,
 "information_density": 3,
 "comments": "The report accurately summarizes
     the invention and controversy of Hawaiian
     pizza, with reliable citations and clear
     attribution of gaps. Coverage is strong
     for the main facts, but lacks granular
     details and direct quotes that would
     improve information density and
     completeness. Citation quality is good,
     but could be improved by linking claims to
      more specific segments. The style is
     concise and factual, but could be more
     engaging with anecdotes. Prioritization is
      appropriate, focusing on invention,
     context, and controversy. Completeness and
      information density are limited by the
     absence of operational details and broader
      cultural context."
}
</eval>
```