Less is More: Improved Fact-Checking on AVeriTeC with Iterative Question Generation

Lucas M. Fraga[*1], **Marcelo M. Delucis**[*1], **Christian Mattjie**[1], **Pedro Wagner**[1], **Arthur Ibarra**[1], **Yasmin Cardozo Aguirre**[1], **Rodrigo C. Barros**[1,2], **Lucas S. Kupssinskü**[1]

[1]MALTA - Machine Learning Theory and Applications Lab
School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, 90619-900, Porto Alegre, RS, Brazil

[2]Kunumi Institute, Brazil

lucas.fraga25@edu.pucrs.br

Abstract. *Automated open-web fact-checking decomposes a claim into questions, retrieves evidence, and produces a verdict.* AVERITEC *systems report gains from iterative question generation, but iteration is bundled with reranking, multi-agent coordination, or fine-tuned generators, leaving its contribution unquantified. We isolate iteration as the only variable in an otherwise fixed* AVERITEC *pipeline and compare it to single-pass generation across six instruction-tuned LLMs (*4-8*B). On the dev set, iterative generation improves accuracy in every case (*1.8 *to* 21.2 *pp) while more than halving queries per claim (*$4.96 \rightarrow 2.38$*) and lowering abstention (*$0.424 \rightarrow 0.221$*). Iteration alone accounts for the accuracy and decisiveness gains reported in prior* AVERITEC *pipelines.*

1. Introduction

The spread of web misinformation has made automated fact-checking a key challenge in NLP, where claims must be verified against evidence retrieved from open sources rather than a fixed corpus. Classification strategies for AFC range from zero- and few-shot prompting of LLMs to task-specific fine-tuning of encoder models, with fine-tuning and evidence grounding proving most effective for veracity prediction [Delucis et al. 2025, Barcelos et al. 2026]. AVERITEC [Schlichtkrull et al. 2023] frames this as claim verification against open-web evidence, providing the dataset subsequent RAG pipelines build on: each claim is decomposed into verification questions, issued as search queries, and the retrieved evidence is aggregated into a verdict, with retrieval quality and grounding driving most reported gains [Trivedi et al. 2023]. Since retrieval runs over the open web, question generation conditions all downstream retrieval and the evidence available at verdict time.

2. Related Work

Within this paradigm, iterative question generation reportedly outperforms single-pass variants, but always alongside other modifications: hybrid Seq2Seq/LLM generators with paraphrasing, verdict reconsideration, and metadata injection [Malon 2024]; multimodal extraction with tree-structured search [Khaliq et al. 2024]; multi-round QA

[*]Equal contribution.

with claim decomposition [Liu et al. 2024b]; multi-agent role specialization and tool use [Hong et al. 2025, Dharmavaram et al. 2025, Gong et al. 2026]; or graph-structured prompting with separately fine-tuned Question-Generation (QG), Question-Answering (QA), and verdict modules [Rolinger et al. 2025].

High-ranking shared-task systems similarly bundle Hypothetical Document Embeddings-style (HyDE) [Gao et al. 2023] query expansion, dense retrieval, and fine-tuned generators [Yoon et al. 2024, Yoon et al. 2025, Liu et al. 2024a]. These fact-checking systems draw on a broader family of iterative-retrieval frameworks that interleave reasoning and acting [Yao et al. 2022], couple retrieval with chain-of-thought [Trivedi et al. 2023], trigger retrieval on token-level uncertainty [Jiang et al. 2023], or fine-tune reflection tokens for adaptive retrieval [Asai et al. 2023]. While these methods iterate during generation rather than during question formulation, they motivate the design choices that AVERITEC pipelines later adopted at the question-generation stage; iteration is one ingredient among several that change simultaneously, and the gains attributable to iteration alone remain unmeasured.

Inconsistent benchmarking has motivated standardized re-implementations on shared datasets [Rau et al. 2024, Jin et al. 2025] and one-factor-at-a-time ablations across the full RAG stack [Wang et al. 2024]. Results show that minimal baselines can match or exceed agentic systems on long-context QA [Laitenberger et al. 2025]; that retrieval follows an inverted-U in the number of chunks, so a single design choice can flip qualitative conclusions [Yu et al. 2024]; and that LM rerankers, standard in AVERITEC pipelines, often fail to beat BM25 [Hagström et al. 2025]. Claims about specific RAG components require controlled comparisons, not aggregate gains over a baseline.

This work isolates iterative question generation as the primary variable in an otherwise fixed AVERITEC pipeline. The distinction is solely whether questions are generated in a single pass, exclusively from the claim, or sequentially, with each question conditioned on the evidence retrieved so far. We run this comparison on six instruction-tuned Large Language Models (LLMs) of comparable scale (DeepSeek-R1:8B [Guo et al. 2025a], Gemma 3:4B and Gemma 4:E4B [Gemma Team et al. 2024], LLaMA 3.1:8B [Grattafiori et al. 2024], Qwen 3:8B [Yang et al. 2025], Ministral 3:8B [Liu et al. 2026]) on the AVERITEC development set, making any resulting difference attributable to iteration alone. Beyond accuracy and F1 score, we measure how iteration affects query budget, abstention rate, and error rate in non-NEI predictions.

3. Experiments

3.1. Dataset

We evaluate our approach on AVERITEC, a dataset of real-world claims requiring web evidence retrieval. Unlike closed-domain datasets, it provides no fixed evidence corpus, making it suitable for evaluating web retrieval-based systems. We use the official dev set of 500 instances, each annotated with one of four labels: SUPPORTED, REFUTED, NOT ENOUGH EVIDENCE, or CONFLICTING EVIDENCE/CHERRY-PICKING. Following prior fact-checking work [Thorne et al. 2018], we refer to NOT ENOUGH EVIDENCE as NEI (NOT ENOUGH INFORMATION).

3.2. Experimental Setup

We propose a controlled comparison between two fact-checking pipelines that differ only in their question generation strategy: a `single-step` approach and an `iterative` approach. Both experiments follow the AVERITEC fact-checking methodology, where a claim is first decomposed into verification questions, which are then used as queries to retrieve relevant evidence from external sources. All code is available in our repository[*].

Task Definition Given a claim c, the goal is to predict a label y as SUPPORTED, REFUTED, NOT ENOUGH EVIDENCE, or CONFLICTING EVIDENCE based on evidence retrieved from the web. The system operates by generating questions to serve as search queries, retrieving relevant documents, and producing a final verdict using an LLM.

Retrieval and Ranking Both approaches use the same retrieval pipeline. Given a query q, we retrieve up to 30 candidate URLs. Documents are then ranked using BM25 and the top 10 candidates are selected. These are split into 300-token passages and passed to a cross-encoder reranker, which selects the final 5 passages used as evidence for that query. The same per-query procedure runs in both algorithms; the conditions differ in the number of queries issued and in how each query is generated. Per-query evidence sets are then accumulated into the pool set $D = \{d_1, \ldots, d_k\}$ used by the verdict step.

Single-Step Question Generation In this approach, the model generates up to a maximum of $N = 5$ queries $\{q_1, \ldots, q_N\}$ given the input claim c. The model is allowed to terminate generation before reaching N if it judges that the questions produced so far are sufficient to support a verdict. Because this decision is made without access to retrieved evidence and relies solely on the model's prior knowledge of the claim, early stopping rarely fires in practice (see Section 4.2 for the average query count under this condition). We retain the mechanism so that both pipelines have a comparable stopping interface and any difference in query budget arises from how the stopping signal is informed, not from whether the option to stop exists.

All generated queries are issued independently to the retrieval pipeline, and the resulting documents are aggregated into a single evidence set D. The LLM then predicts y from the claim c and aggregated evidence D. Algorithm 1 gives the full procedure.

Algorithm 1 Single-Step Question Generation

Require: claim c, language model f_{LLM}, retrieval function RETRIEVE, maximum questions $N = 5$
Ensure: verdict y
$\{q_1, \ldots, q_n\} \leftarrow f_{\text{LLM}}(c)$ ▷ generate up to N questions in one pass, $n \leq N$
$\mathcal{D} \leftarrow \emptyset$ ▷ accumulated evidence
for $i = 1$ **to** n **do**
    $D_i \leftarrow \text{RETRIEVE}(q_i)$, $\mathcal{D} \leftarrow \mathcal{D} \cup D_i$
end for
$y \leftarrow f_{\text{LLM}}(c, \mathcal{D})$ ▷ produce final verdict from aggregated evidence
return y

Iterative Question Generation In contrast, the iterative approach generates questions sequentially, conditioning each on previously retrieved evidence. Starting from the claim c, the model generates a question q_t at each iteration t; after retrieving evidence D_t, it decides whether the available information suffices for a final prediction or whether further

[*] `https://github.com/Malta-Lab/iterative-question-generation`

questions are needed. Iterations continue until the model outputs a final label or reaches the maximum of 5 iterations.

We cap t at 5 for computational budget: each iteration requires retrieval, reranking, and LLM call, and a larger budget would make evaluation infeasible. For fair comparison, the number of questions N in single-step generation is also limited to 5. This matches the query count in prior AVERITEC submissions, suggesting that 5 is a reasonable operating point rather than a restrictive one. Algorithm 2 gives the full procedure.

Algorithm 2 Iterative Question Generation

Require: claim c, language model f_{LLM}, retrieval function RETRIEVE, maximum iterations $T = 5$
Ensure: verdict y

$\mathcal{Q} \leftarrow \emptyset$ ▷ accumulated queries
$\mathcal{D} \leftarrow \emptyset$ ▷ accumulated evidence
for $t = 1$ **to** T **do**
  $q_t, \texttt{stop} \leftarrow f_{\text{LLM}}(c, \mathcal{Q}, \mathcal{D})$ ▷ generate query or signal stop
  if `stop` **then**
    break
  end if
  $D_t \leftarrow \text{RETRIEVE}(q_t)$
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q_t\}, \mathcal{D} \leftarrow \mathcal{D} \cup D_t$
end for
$y \leftarrow f_{\text{LLM}}(c, \mathcal{D})$ ▷ produce final verdict from accumulated evidence
return y

At each step, the model is conditioned on the claim, all previously generated questions, and all previously retrieved evidence. It then either issues a new question or signals that the accumulated evidence is sufficient, in which case the loop terminates. This design ensures a controlled comparison that isolates the effect of iterative conditioning in question generation. Because the iterative pipeline issues fewer queries and stops earlier, it also accumulates fewer retrieved passages at verdict time than our single-step pipeline. We attribute the resulting performance differences jointly to better-targeted questions and to a more focused evidence pool, which are coupled effects of the same conditioning.

3.3. Models

We evaluate six instruction-tuned LLMs in the $4 - 8$B parameter range: DeepSeek-R1:8B, Gemma 3:4B and Gemma 4:E4B, LLaMA 3.1:8B, Ministral 3:8B, and Qwen 3:8B. The Gemma 4:E4B model uses Per-Layer Embeddings, yielding 4.5B effective parameters (8B total with embeddings). This selection spans 3 architectural families and 2 training paradigms (standard instruction tuning and reasoning- distilled), without tuning model selection to outcomes. Within this range, recent work suggests that small LLMs can achieve competitive performance in reasoning and RAG settings while remaining locally executable [Nissen et al. 2025, Shah et al. 2026]. LLaMA 3.1 serves as a stable general-purpose reference [Grattafiori et al. 2024]; DeepSeek-R1 represents reasoning-distilled training [Guo et al. 2025a]; Qwen 3 and Ministral 3 cover instruction-following and structured-generation profiles [Yang et al. 2025, Liu et al. 2026]; and Gemma 3:4B and Gemma 4:E4B add architectures targeting performance-to-size and on-device efficiency [Gemma Team et al. 2024, Bai et al. 2024]. Whether iterative question generation benefits from specific model capabilities or transfers across these profiles is an empirical question that our cross-model evaluation directly addresses [Guo et al. 2025b]. Models are run locally with temperature set to 0.0 and standard configuration to ensure deterministic outputs. The maximum number of generated questions is limited to 5, while the stopping criteria are determined by the model itself [Willard et al. 2023].

3.4. Web Search Implementation Details

We adopt DuckDuckGo [*] as an alternative to commercial search APIs [Putta et al. 2025, Rashid et al. 2025], trading result stability for accessibility. Unlike commercial engines, it guarantees neither stable rankings nor reliable control over document timestamps, as its ranking often prioritizes recent or paid content [Hasanain et al. 2022]. Rather than enforcing strict temporal constraints, we adopt a *retrieval-first* approach prioritizing evidence coverage under realistic web conditions, a trade-off between realism and experimental control that remains an open challenge [Nakov et al. 2021, Chen et al. 2024]. Retrieved documents are split into 300-token passages and reranked with `ms-marco-MiniLM-L-6-v2`[*], taking the top-5 as evidence. To mitigate the reproducibility risk this approach introduces, we cache every retrieved page in our public repository and make available all the records of the experiments conducted, including the evidence set used at each step of the reported experiments, thereby allowing our results to be audited and replicated independently of future DuckDuckGo query outcomes.

4. Results & Discussion

4.1. Overall Performance

Table 1 presents the overall performance of single-step and iterative experiments across all six evaluated models. Iterative generation improves accuracy in all cases, with gains ranging from 1.8 up to 21.2 percentage points. The largest improvements are observed for Ministral3 (+21.2 pp), Qwen3 (+16.8 pp), and DeepSeek-R1 (+14.0 pp), all of which had relatively weak single-step baselines. Gemma3's single-step baseline is the strongest among all evaluated models (0.598 accuracy), and iterative generation yields a marginal gain (+1.8 pp). Notably, its macro-F1 score decreases slightly under iterative generation ($0.342 \rightarrow 0.337$), driven by a collapse in NEI recall, as analyzed in Section 4.3.

Accuracy gains are consistent across the SUPPORTED and REFUTED labels, as reflected in per-class F1. These are the classes where web evidence is most informative, and iterative retrieval improves coverage precisely where it matters most. NEI performance stays low across all models and conditions, indicating that identifying insufficient evidence remains an open challenge regardless of question generation strategy.

4.2. Efficiency Trade-offs

To characterize token-level cost, we report the average number of queries and token consumption per claim in Table 2. The results indicate a reallocation of computational cost: iterative generation reduces the average number of queries by more than half ($4.96 \rightarrow 2.38$) while increasing average token consumption ($14{,}624 \rightarrow 16{,}681$).

The reduction in queries follows directly from the iterative stopping criterion: at each step, the model evaluates whether the accumulated evidence suffices to issue a verdict. Conditioning on previously retrieved evidence enables earlier convergence and avoids the defensive over-generation seen in single-step pipelines, where the model issues up to five questions without intermediate feedback. The corresponding rise in token consumption reflects the longer, evidence-conditioned prompts produced at each step.

[*] `https://github.com/deedy5/ddgs`

[*] `https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2`

Table 1. Comparison between single-step and iterative question generation. Iterative generation consistently improves performance. F1 refers to macro-averaged F1. Class-wise F1 highlights strong results on SUP and REF, while NEI remains challenging.

Model	Method	Acc	F1	$F1_{SUP}$	$F1_{REF}$	$F1_{NEI}$
DeepSeek-R1:8B	Single-step	0.424	0.405	0.433	0.605	**0.176**
	Iterative	**0.564**	**0.477**	**0.543**	**0.735**	0.153
Gemma3:4B	Single-step	0.598	**0.342**	0.242	0.753	**0.031**
	Iterative	**0.616**	0.337	**0.255**	**0.756**	0.000
Gemma4-E4B	Single-step	0.372	0.369	0.382	0.556	**0.171**
	Iterative	**0.480**	**0.451**	**0.569**	**0.647**	0.136
LLaMA3.1:8B	Single-step	0.502	0.405	0.398	0.678	0.141
	Iterative	**0.608**	**0.480**	**0.512**	**0.740**	**0.187**
Ministral3:8B	Single-step	0.402	0.336	0.267	0.595	0.148
	Iterative	**0.614**	**0.454**	**0.398**	**0.781**	**0.183**
Qwen3:8B	Single-step	0.378	0.380	0.426	0.543	0.172
	Iterative	**0.546**	**0.496**	**0.622**	**0.693**	**0.174**
Average	Single-step	0.446	0.359	0.351	0.589	**0.140**
	Iterative	**0.538**	**0.426**	**0.463**	**0.677**	0.139

From a practical standpoint, this trade-off favors iterative generation in settings where retrieval is the primary bottleneck, such as pipelines that rely on paid search APIs. In compute-constrained environments, the increased token usage may warrant consideration, though the accuracy gains reported in Section 4.1 suggest the overhead is well justified.

4.3. Reduction in Abstention

Table 3 shows the proportion of NEI predictions per model. Iterative generation consistently reduces abstention across all six models, with the average NEI rate dropping from 0.424 to 0.221. The largest reductions occur for Ministral3 (-0.302), Qwen3 (-0.286), and DeepSeek-R1 (-0.250), the same models with the largest accuracy gains in Section 4.1. These findings suggest that reduced abstention is a key driver of overall improvement.

By conditioning each query on previously retrieved evidence, the model is able to refine its search iteratively and produce a verdict more often. In the single-step setting, the model issues all queries without feedback and must decide based on whatever evidence is found, a noisier process that often yields insufficient grounding for a confident verdict.

Gemma3 is again an outlier, this time in the opposite direction: its single-step NEI rate is already very low (0.060), indicating it rarely abstains regardless of available evidence. Iterative generation cuts abstention to 0.022, effectively eliminating it and, as noted in Section 4.1, collapsing its NEI F1 to zero. This suggests Gemma3's behavior is driven less by evidence quality than by a strong prior toward decisive predictions.

Table 2. Efficiency comparison between single-step (S) and iterative (I) question generation.

Model	Queries (S)	Queries (I)	Tokens (S → I)
DeepSeek-R1:8B	4.87	2.31	15636 → 16518
Gemma3:4B	4.98	2.34	15463 → 16655
Gemma4-E4B	4.09	2.39	12590 → 16427
LLaMA3.1:8B	4.89	2.35	15355 → 17352
Ministral3:8B	5.00	2.51	13322 → 16856
Qwen3:8B	4.92	2.37	15381 → 17279
Average	4.96	2.38	14624 → 16681

Table 3. Reduction in abstention (NEI). Iterative generation consistently decreases uncertainty across all models, leading to more decisive predictions.

Model	Single	Iterative	Δ
DeepSeek-R1:8B	0.520	0.270	**-0.250**
Gemma3:4B	0.060	0.022	**-0.038**
Gemma4-E4B	0.586	0.400	**-0.186**
LLaMA3.1:8B	0.300	0.144	**-0.156**
Ministral3:8B	0.472	0.170	**-0.302**
Qwen3:8B	0.606	0.320	**-0.286**
Average	0.424	0.221	**-0.203**

4.4. Reliability of Predictions

Nevertheless, reducing abstention does not guarantee improved reliability; a model that abstains less may simply be making more mistakes. We report the error rate conditioned on non-NEI verdicts to assess this trade-off, presented in Table 4. For LLaMA3.1 and Ministral3, iterative generation reduces both abstention and error rate, indicating genuine improvements in coverage and precision. For DeepSeek-R1 and Qwen3, the error rate rises slightly ($+0.038$ and $+0.059$, respectively), suggesting that some of the additional decisions made under iterative generation are incorrect. Gemma3 and Gemma4 show relatively stable error rates, with changes of less than 0.02 in both cases.

Despite this increase in non-NEI errors for some models, overall accuracy improves across all six (Section 4.1), indicating that the benefit of committing to a verdict, even at the cost of some additional errors, outweighs the cost of abstaining on verifiable claims. Iterative generation shifts the system toward greater decisiveness, and in most cases the evidence gathered through iterative retrieval is sufficient to justify it.

We illustrate this behavior for DeepSeek-R1:8B in Figure 1. Under single-step generation, 58% of truly SUPPORTED claims are predicted as NEI: the model frequently fails to gather enough evidence to confirm a claim. Iterative generation reduces this to 30%, with a corresponding rise in correct SUPPORTED predictions ($0.30 \rightarrow 0.44$), REFUTED claims follow with $0.49 \rightarrow 0.70$.

Table 4. Error rate conditioned on non-NEI predictions. While iterative generation reduces abstention, its impact on reliability varies across models, revealing a trade-off between decisiveness and accuracy.

Model	Single	Iterative	Δ
DeepSeek-R1:8B	0.225	0.263	**+0.038**
Gemma3:4B	0.366	0.370	**+0.004**
Gemma4-E4B	0.237	0.253	**+0.017**
LLaMA3.1:8B	0.320	0.313	**-0.007**
Ministral3:8B	0.314	0.287	**-0.028**
Qwen3:8B	0.188	0.247	**+0.059**
Average	0.293	0.319	**+0.026**

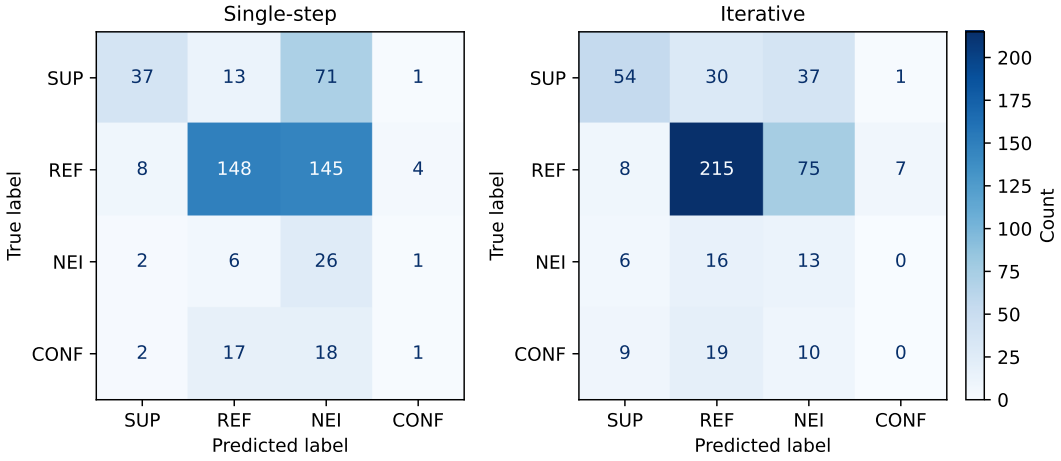


Figure 1. Confusion matrices of DeepSeek-R1:8B.

Two further patterns appear in the confusion matrices. First, when iterative generation causes the model to abandon NEI verdicts on truly NEI claims, the errors are not evenly distributed: 46% of true NEI claims are misclassified as REFUTED, suggesting a bias toward negative verdicts when evidence is ambiguous but present. Second, predictions of the CONFLICTING EVIDENCE class collapse entirely under iterative generation ($0.03 \rightarrow 0.00$): the model stops hedging and commits to either SUPPORTED or REFUTED. This is consistent with the trend of increased decisiveness; as evidence accumulates, the model resolves ambiguity by taking a side rather than acknowledging conflict.

This decisiveness may be compounded by temporal drift in evidence: AVERITEC labels were assigned in 2023, so claims annotated CONFLICTING EVIDENCE or NOT ENOUGH EVIDENCE then may now be resolvable as more sources have become available. By issuing more targeted queries, iterative retrieval may be more likely to surface these evidence. Figure 2 presents an example: for a claim about Sean Connery and Apple, single-step generation produces 5 broad questions and returns NOT ENOUGH EVIDENCE, while iterative generation converges in 3 targeted queries and returns a REFUTED verdict.

4.5. Limitations

While our findings are consistent across all six models, our experimental scope is controlled along four dimensions. Regarding the extent of retrieval, DuckDuckGo does not guarantee stable rankings or consistent document availability over time, which may introduce variability if experiments are repeated at different times. On the benchmark spectrum, all experiments are conducted on the AVERITEC dev set (500 instances), and whether the observed gains generalize to other fact-checking datasets or claim domains remains an open question. On the model domain, all evaluated LLMs are small-scale (up to 8B parameters), locally run at temperature 0.0, and the extent to which findings hold for larger models, proprietary APIs, or stochastic settings has not been assessed. Finally, AVERITEC labels were defined in 2023 and may be outdated, especially regarding classes CONFLICTING EVIDENCE and NOT ENOUGH EVIDENCE. Further evidence available today may justify some of these claims to be considered as SUPPORTED or REFUTED.

Claim Ground truth: **REFUTED**

In a letter to Steve Jobs, Sean Connery refused to appear in an Apple commercial.

Single-step question generation **✗ NOT ENOUGH EVIDENCE**

1. Was Sean Connery **alive and professionally active** when Apple sought commercial actors?
2. Did Steve Jobs ever **write a letter** to Sean Connery about an advertisement?
3. Is there any **record or documentation** of Sean Connery refusing an Apple commercial?
4. Was Steve Jobs still **alive and working at Apple** at that time?
5. Has anyone **publicly confirmed or denied** such an event?

Iterative question generation **✓ REFUTED**

1. Is there any **credible source or evidence** supporting the alleged letter?
2. Is the claim based on a **satirical or unreliable source**?
3. Was there any **real report of Apple approaching** Sean Connery for a commercial?

Figure 2. Single-step vs. iterative question generation in LLaMA 3.1:8B.

5. Conclusion and Future Work

Prior work on iterative question generation for fact-checking bundles iteration with other pipeline modifications: reranking, multi-agent coordination, fine-tuned generators, which makes it hard to attribute gains to iteration alone. We isolate it as the only variable in an otherwise fixed AVERITEC pipeline. Across six instruction-tuned LLMs, iterative generation improves verification accuracy in every case (1.8 to 21.2 pp) while more than halving the average questions per claim (4.96 $\rightarrow$ 2.38).

The mechanism is a shift in how the system handles uncertainty: when each question is conditioned on previously retrieved evidence, the model abstains less (NEI rate 0.424 $\rightarrow$ 0.221) and commits to more decisive verdicts. Iterative generation does raise token consumption and, for some models, slightly increases the error rate among non-NEI verdicts, but these setbacks are minor against the accuracy gains.

Several directions remain open. First, evaluating iterative generation beyond AVERITEC would clarify whether the gains generalize across claim domains and retrieval

settings. Second, scaling to larger models would show whether the benefits persist as capacity increases. Third, with iteration's contribution now isolated, future work can reintroduce other pipeline components in a controlled manner, measuring their incremental gains over a well-characterized iterative baseline.

Acknowledgements This study was financed in part by the Coordination for the Improvement of Higher Education Personnel (CAPES) — Finance Code 001; by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) — Grant Number: 443072/2024-8; and by Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) — Grant Number: 25/2551-0000891-3 and 25/2551-0002824-8. This paper was achieved in cooperation with HP Brasil Indústria e Comércio de Equipamentos Eletrônicos LTDA. using incentives of Brazilian Informatics Law (Law nº 8.2.48 of 1991). This work was supported by Kunumi Institute. The authors thank the institution for its financial support and commitment to advancing scientific research.

References

Asai, A. et al. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Bai, G. et al. (2024). Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*.

Barcelos, P., Parraga, O., Mussi, M. M., Fraga, L. M., Kupssinskü, L. S., and Barros, R. C. (2026). Latent fact-checking: Detecting misinformation through activation engineering.

Chen, J. et al. (2024). Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587.

Delucis, M. et al. (2025). Automated fact-checking in brazilian portuguese: Resources and baselines. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 137–148. SBC.

Dharmavaram, A. et al. (2025). SANCTUARY: An efficient evidence-based automated fact checking system. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 247–257, Vienna, Austria. Association for Computational Linguistics.

Gao, L. et al. (2023). Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Gemma Team, Mesnard, T., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gong, S. et al. (2026). Multi-sourced, multi-agent evidence retrieval for fact-checking. *arXiv preprint arXiv:2603.00267*.

Grattafiori, A. et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, D. et al. (2025a). Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Guo, J. et al. (2025b). Stylebench: Evaluating thinking styles in large language models. *arXiv preprint arXiv:2509.20868*.

Hagström, L. et al. (2025). Language model re-rankers are fooled by lexical similarities. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 18–33, Vienna, Austria. Association for Computational Linguistics.

Hasanain, M. et al. (2022). Studying effectiveness of web search for fact checking. *J. Assoc. Inf. Sci. Technol.*, 73(5):738–751.

Hong, S. et al. (2025). EMULATE: A multi-agent framework for determining the veracity of atomic claims by emulating human actions. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 179–183, Vienna, Austria. Association for Computational Linguistics.

Jiang, Z. et al. (2023). Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Jin, J. et al. (2025). Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In *Companion Proceedings of the ACM on Web Conference 2025*, page 737–740, New York, NY, USA. Association for Computing Machinery.

Khaliq, M. A. et al. (2024). RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Laitenberger, A. et al. (2025). Stronger baselines for retrieval-augmented generation with long-context language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32559–32569, Suzhou, China. Association for Computational Linguistics.

Liu, A. H. et al. (2026). Ministral 3. *arXiv preprint arXiv:2601.08584*.

Liu, J. et al. (2024a). FZI-WIM at AVeriTeC shared task: Real-world fact-checking with question answering. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 77–85, Miami, Florida, USA. Association for Computational Linguistics.

Liu, J. et al. (2024b). GProofT: A multi-dimension multi-round fact checking framework based on claim fact extraction. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 118–129, Miami, Florida, USA. Association for Computational Linguistics.

Malon, C. (2024). Multi-hop evidence pursuit meets the web: Team papelo at FEVER 2024. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 27–36. Association for Computational Linguistics.

Nakov, P. et al. (2021). Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

Nissen, L. et al. (2025). Medicine on the edge: Comparative performance analysis of on-device llms for clinical reasoning.

Putta, A. R. et al. (2025). ClaimCheck: Automatic fact-checking of textual claims using web evidence. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 303–316, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Rashid, F. B. et al. (2025). Fathom: A fast and modular RAG pipeline for fact-checking. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 258–265, Vienna, Austria. Association for Computational Linguistics.

Rau, D. et al. (2024). BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, Miami, Florida, USA. Association for Computational Linguistics.

Rolinger, S. et al. (2025). Graph-of-thoughts for fact-checking with large language models. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 266–273, Vienna, Austria. Association for Computational Linguistics.

Schlichtkrull, M. et al. (2023). Averitec: A dataset for real-world claim verification with evidence from the web.

Shah, A. P. et al. (2026). Small language models: Architecture, evolution, and the future of artificial intelligence.

Thorne, J. et al. (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.

Trivedi, H. et al. (2023). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Wang, X. et al. (2024). Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736, Miami, Florida, USA. Association for Computational Linguistics.

Willard, B. T. et al. (2023). Efficient guided generation for large language models.

Yang, A. et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yao, S. et al. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yoon, Y. et al. (2024). Hero at averitec: The herd of open large language models for verifying real-world claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 130–136.

Yoon, Y. et al. (2025). Team HUMANE at AVeriTeC 2025: HerO 2 for efficient fact verification. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 224–228, Vienna, Austria. Association for Computational Linguistics.

Yu, T. et al. (2024). In defense of rag in the era of long-context language models. *arXiv preprint arXiv:2409.01666*.