CLEAR: Contamination-Aware Evaluation of Retrieval in Agentic Fact-Checking

Anonymous ACL submission

Abstract

Evaluating information retrieval in agentic systems is increasingly difficult due to model contamination and tight coupling between retrieval and intervened agent reasoning. Large language models may recall fact checking knowledge from pretraining, while agents shape queries in ways that confound retrieval evaluation, causing standard end to end evaluations to yield conclusions that do not generalize across agentic architectures or datasets. We introduce a contamination aware evaluation framework for retrieval in agentic fact checking that fixes the language model and corpus and evaluates retrieval across diverse agentic-retriever interaction settings, enabling controlled analysis of how contamination and query generation affect retrieval quality independently of downstream reasoning. Our experiments show that contamination impacts retrieval behavior, retriever rankings are unstable across agentic systems due to query and retrieval interaction effects, and that different choices of how NDCG values are aggregated can lead to qualitatively different and even reversed comparisons between agents. For datasets with silver documents, we propose nDEv2R, a rank sensitive fact level retrieval metric that remains informative under incomplete evidence supervision. While instantiated in fact checking, our findings apply more broadly to evaluating retrieval components embedded in agentic systems such as question answering and multi document reasoning.

1 Introduction

Misinformation undermines trust in democratic processes, science, and institutions, contributing to poor decision-making with harmful real-world consequences (Zarocostas, 2020). Large language models (LLMs) amplify this problem through hallucinations or deliberate misuse to generate disinformation at scale (Augenstein et al., 2024; Wang et al., 2025). Since the volume of misinformation now exceeds human fact-checking capacity, automated fact-checking (AFC) has become essential (Vlachos and Riedel, 2014). Following journalistic practices, AFC systems retrieve relevant evidence and ground their verdicts in it (Guo et al., 2022). This evidence-based approach enhances transparency, trust, and accountability, critical factors for responsible real-world deployment (Nakov et al., 2021; Warren et al., 2025).

Despite the central role of evidence collection in both journalistic and automated fact-checking (Arnold, 2020), retrieval methods for AFC remain understudied. Recent AFC architectures use agentic designs that decompose claims into subclaims or questions and retrieve evidence using drop-in search engines, often commercial ones (Xie et al., 2025; Braun et al., 2025; Vladika et al., 2025). While such designs improve evidence coverage, they complicate evaluation: retrieval performance becomes entangled with the agent's reasoning and decomposition strategies, making fair comparisons across methods nearly impossible. Data contamination poses an additional challenge. LLMs may memorize information from pretraining (Sainz et al., 2023), allowing them to recall why claims are incorrect and generate useful subqueries or subquestions while bypassing the typical difficulties of formulating queries from novel claims alone to find evidence (Glockner et al., 2022). Commercial search engines exacerbate this issue by prioritizing verified fact-checking articles (Koronska and Rogers, 2024), further simplifying the task for previously fact-checked claims. Moreover, dependence on commercial search engines limits the applicability of fact-checking systems in domains involving sensitive or non-public data, as is common in enterprise settings (Bruckhaus, 2024). Disentangling retrieval methods from modern agentic systems is therefore an important yet understudied research direction.

We present CLEAR (Contamination-aware

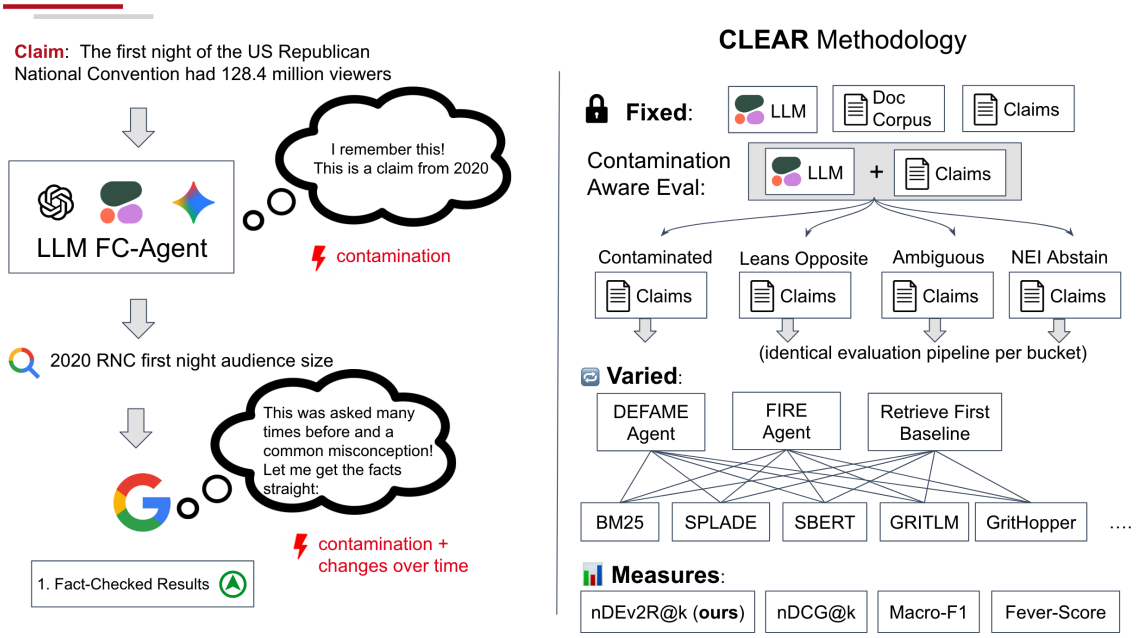

Figure 1: Overview of the CLEAR methodology. **Left:** In realistic fact-checking settings, contamination can arise from both language models and external search engines: LLMs may recall claims from pretraining, while search results can change over time and favor previously fact-checked sources. **Right:** CLEAR enables controlled evaluation by fixing the LLM, document corpus, and claims, and stratifying claims via LLM-only probing into contamination categories. Each category is evaluated using an identical pipeline while varying agent architectures and retrieval methods, allowing analysis of agent-retriever interaction effects using standard metrics and our proposed nDEv2R@k.

Longitudinal Evaluation of Agentic Retrieval), a novel contamination-aware evaluation methodology comprising 3 diverse single and multi-hop fact-checking datasets for evaluating the interaction between retrieval methods and agent-based fact-checking systems. We design CLEAR to examine how retrieval method choices across different agentic *architectures* affect the performance in (i) selecting the correct verdict, (ii) identifying sufficient evidence, and (iii) achieving both jointly. To isolate the effects of retrieval and system architecture, CLEAR is built around a fixed LLM that remains constant across all experiments. We opt for this design choice since it enables controlled experiments regarding data contamination and its impact on claim decomposition and evidence retrieval. Building on CLEAR, we investigate the following two research questions:

- **RQ1:** Do contamination effects impact retrieval performance on fact-checking datasets?
- **RQ2:** Is the performance of the evidence retrieval method independent of the agentic system for unseen claims?

Our experiments on sparse, dense and multi-hop retrieval method families show that agentic systems perform substantially better at identifying relevant evidence when the veracity label is known, suggesting that they unfairly benefit from knowing how to decompose claims, and that data contamination affects not only veracity prediction but also confounds evidence retrieval. We further observe that the performance of retrieval methods varies substantially across agentic systems. The same retriever can behave very differently depending on the system in which it is deployed. These strong dependencies hinder the independent development and evaluation of new retrieval techniques and confound performance attribution, making it difficult to isolate the impact of specific design choices. To improve this, CLEAR can be easily extended with additional agentic systems, facilitating the development of robust retrieval methods that are decoupled from any particular agentic system implementation.

2 CLEAR: Contamination-aware Longitudinal Evaluation of Agentic Retrieval

We design the our method with four principles: (i) end-to-end evaluation with retrieval in-the-loop;

(ii) modularity to swap retrievers and AFC architectures; (iii) reproducibility via fixed corpora and controlled retrieval; and (iv) coverage across single- and multi-hop claims.

2.1 CLEAR Method

The goal of CLEAR is to evaluate the performance of retrieval methods used by agentic fact-checking systems. We define an agentic fact-checking system as any LLM-based system that collects evidence documents and reasons over them to produce a veracity label. The agent may either pass the original claim directly to the retrieval component, as in traditional fact-checking pipelines (Thorne et al., 2018; DeHaven and Scott, 2023), or first decompose claims into sub-claims or questions (Ousidhoum et al., 2022; Chen et al., 2024a). We define the evidence retrieval task as follows: Given a claim c and a retrieval method R integrated into an agentic fact-checking system F, the retrieval method R must identify the correct and complete set of evidence documents to enable the fact-checking system F to make the veracity prediction. Our task definition is intentionally conditioned on a fixed set of fact-checking systems with a fixed LLM, allowing us to systematically evaluate the retrieval method's capabilities across diverse yet representative claim decomposition strategies when integrated into an agentic system, while accounting for potential data contamination when the LLM has previously encountered information about a claim during pretraining (Sainz et al., 2023).

Isolating the impact of individual design choices in modern AFC systems is intractable. We therefore operate under two assumptions: First, the AFC system is not tied to a specific retrieval method, allowing retrieval components to be replaced and evaluated independently (*retrieval independence*). Second, the AFC system is not tied to a specific LLM, enabling systematic analysis of contamination effects by fixing the underlying model (*LLM independence*). Both assumptions should hold for general AFC systems that support independent incorporation and improvement of retrieval and LLM components, which is the focus of this study.

2.2 CLEAR Instantiation

We instantiate the CLEAR using three Fact-Checking agents, three datasets and in total seven retrievers which we outline in the following sections. To study contamination effects in a controlled setting (RQ1), we fix the underlying language model across all experiments, as model-side contamination depends directly on the LLM's pretraining and memorized knowledge. We use `Command-A` from Cohere (Cohere et al., 2025) for all agents and datasets. We discuss hyperparameters in Appendix D. Our code is anonymized and publicly available at `https://anonymous.4open.science/r/CLEAR-B03C`.

2.3 Datasets

We select three text-based fact-checking datasets that cover complementary retrieval regimes relevant to agentic fact-checking. **SciFactOpen** targets high-precision retrieval for scientific claims using evidence from research abstracts (Wadden et al., 2022). **AveriTeC** contains real-world claims from news and the web with long, heterogeneous documents and silver, incomplete evidence supervision (Schlichtkrull et al., 2023). **Ex-FEVER** provides a controlled multi-hop setting over Wikipedia, where claims require aggregating evidence across multiple documents (Ma et al., 2024). Together, these datasets span single- and multi-hop retrieval, gold and silver supervision, and diverse corpus characteristics. Dataset-specific preprocessing and adaptations to our retrieval-centric setup are described in Appendix A.

2.4 Contamination Analysis

For every claim we compare the sampled verdicts to the gold label and assign the claim to one of four buckets:

- **Contaminated**: the model outputs the gold verdict in at least nine of the ten runs (a strong hint that the claim is memorized).
- **Leans Opposite**: the gold verdict appears zero or one times, indicating the model consistently predicts an incorrect label.
- **Ambiguous**: the gold verdict appears two to eight times; the model is inconsistent, so we treat the claim as uncertain rather than clearly contaminated.
- **NEI-Abstain**: the model answers NOT ENOUGH INFO in nine or more runs. Because the model explicitly refuses to commit, we do not flag these cases as contamination even if the gold label is also NEI. Important: Claims whose gold label is NOT ENOUGH INFO are *never* marked contaminated: if the model answers NEI nine or more times we tag them as NEI-abstain; if it confidently predicts a non-NEI label they count as leans-false; otherwise they

remain ambiguous.

Table 1 reports the resulting distribution. The EX-Fever subset shows moderate contamination: 27.0% of the claims are solved perfectly without retrieval, with the largest share (33.4%) falling into the ambiguous category and 27.8% leading to NEI responses. AveriTeC (500 claims) shows a similar pattern: one third contaminated (33.0%), one third NEI responses (33.4%), with ambiguous and leaning-false claims comprising the remainder. SciFact exhibits the highest contamination rate at over one third (35.3%), with substantial ambiguous (30.3%) and NEI-abstain (23.5%) shares, underscoring how the model handles specialized biomedical knowledge with varying degrees of certainty. We use these buckets to stratify subsequent analyses and to measure whether retrieval remains helpful once uncontaminated claims are isolated.

2.5 AFC Architectures

In CLEAR, we evaluate retrieval methods within three representative AFC architectures. As a baseline, we use a vanilla retrieve-first model that directly passes the claim to a retriever and verifies it using the top-k retrieved documents. As agentic AFC systems, we adopt DEFAME (Braun et al., 2025) and FIRE (Xie et al., 2025), two state-of-the-art approaches that iteratively generate subqueries, maintain a running evidence state, and output a veracity label only after sufficient evidence has been collected. Both systems come with their specific design choices and prompts, which we keep as-is to preserve their original behavior and avoid adaptation to specific retrieval methods.

2.6 Retrieval Methods

Following current surveys on retrieval-augmented systems (Chen et al., 2024b; Fan et al., 2024; Gao et al., 2023), retrieval techniques can be broadly grouped into sparse lexical, dense semantic, interaction-aware/multi-vector. To represent these major categories and to cover both frequently used baselines and competitive state-of-the-art methods in each regime, we select a compact set of retrievers for evaluation.

For *sparse lexical* retrieval, we include **BM25** (Robertson and Zaragoza, 2009) as a widely adopted baseline in many RAG and fact-checking pipelines, and **SPLADE** (Formal et al., 2021) as a learned sparse method that extends lexical matching with neural term weighting.

For *dense retrieval*, we evaluate bi-encoder style models that embed queries and documents into a shared space: **Sentence Transformers** (Reimers and Gurevych, 2019) and **Contriever** (Izacard et al., 2022) are commonly used off–the–shelf dense retrievers in NLP benchmarks, while **GritLM** (Muennighoff et al., 2024) reflects recent advances in instruction-tuned dense retrieval showing strong performance across tasks.

To capture methods with richer query–document interaction, we include **ColBERTv2** (Santhanam et al., 2022), a late-interaction retriever that balances semantic richness and efficiency.

Finally, for settings requiring multi-step evidence aggregation, we include **GritHopper** (Erker et al., 2025) as a representative multi-hop retriever that conditions on previously retrieved evidence.

All methods are used off-the-shelf without additional fine-tuning to ensure that differences in performance reflect their general capabilities and how they interact with agentic fact-checking pipelines rather than task-specific adaptation.

2.7 Metrics

Two common approaches exist for evaluating evidence in fact-checking. The first (*reference-based*) assumes knowledge of which documents or information are needed to predict the veracity label (Thorne et al., 2018; Aly et al., 2021; Schlichtkrull et al., 2023), enabling strict evaluation but presuming a single path to the correct label. The second (*veracity-based*) assesses evidence quality indirectly via the veracity prediction, offering more flexibility but unable to detect predictions made with insufficient evidence (Akhtar et al., 2024). In this work, we consider both complementary approaches and report the following metrics. For veracity-based metrics, we report Macro F_1.

Gold references On datasets with complete gold evidence annotations, we evaluate retrieval using nDCG@10 as a reference-based metric. For each claim, we align each gold document to the retrieval iteration where it achieved its best (lowest) rank, then compute sample-level nDCG@10 aggregated across all gold documents. This *gold-document-oriented* aggregation measures whether the agent eventually retrieves relevant evidence, but it favors multi-hop agents that perform many retrieval attempts. To account for retrieval cost, we additionally compute a hop-regularized variant that divides the gold-document-oriented nDCG by the logarithm of the number of retrieval iterations plus

Dataset	Contaminated	Ambiguous	Leans Opposite	NEI-Abstain	Number of Claims
EX-Fever	27.0%	33.4%	11.8%	27.8%	500
AveriTeC	33.0%	18.8%	14.8%	33.4%	500
SciFact	35.3%	30.3%	10.9%	23.5%	279

Table 1: Contamination categories derived from 10 predictions per claim without providing evidence. A claim is marked contaminated when the model reproduces the gold label in at least nine samples; leans false when it matches once or never; ambiguous otherwise; and NEI-Abstain when the model answers NOT ENOUGH INFO in at least nine samples.

one (nDCG@10/ $\log(\text{hops}+1)$), which we find correlates most strongly with veracity prediction. We explore additional aggregation strategies in Appendix C.

Silver references When evidence annotations are incomplete or silver, as in AveriTeC (Schlichtkrull et al., 2023), this assumption no longer holds. Multiple documents may express overlapping or partial factual support, and useful evidence may not be explicitly annotated as gold. In such settings, document-level relevance metrics become brittle, as binary judgments fail to capture degrees of factual support. Ev2R (Akhtar et al., 2025) addresses this limitation by evaluating evidence at the fact level using LLMs, computing a F_1 score based on factual overlap between retrieved documents and reference evidence. While this provides a more informative signal than binary document relevance, Ev2R evaluates documents in isolation and does not account for the ranked nature of retrieval. This is limiting in agentic fact-checking systems, where retrieval outputs are consumed as ordered lists and rank strongly influences downstream usage across different agent architectures .

We therefore propose **nDEv2R@k** (normalized discounted Ev2R), a rank-sensitive extension of Ev2R that evaluates retrieval quality as an ordered list of documents. Let $\mathcal{G} = \{g_1, \ldots, g_m\}$ denote the set of gold evidence documents for a claim, and let d_i be the retrieved document at rank i. We define

$$\text{nDEv2R@k} = \frac{\sum_{i=1}^{k} \frac{\text{Agg}_{g \in \mathcal{G}} \text{Ev2R}(d_i, g)}{\log_2(i+1)}}{\sum_{i=1}^{k} \frac{1}{\log_2(i+1)}},$$

where Agg aggregates fact-level overlap scores across the gold evidence set. In our experiments, we instantiate Agg as either a maximum, reflecting a sufficiency-based assumption where one informative gold document is enough (similar to the Fever-score), or as a mean, reflecting coverage over all annotated gold documents. This unified formulation yields an nDCG-style metric that is rank-sensitive, bounded in $[0, 1]$, and independent of how downstream agents consume retrieved evidence. This formulation yields a rank-sensitive retrieval metric that directly captures how the ordering of retrieved documents affects downstream evidence availability, rather than veracity prediction. In our analysis, we therefore treat nDEv2R@k as a pure retriever-quality measure and contrast it against unweighted Ev2R aggregations to isolate the effect of ranking.

3 Contamination Impact on Retrieval

We begin by analyzing how model-side contamination affects both veracity prediction and evidence retrieval (RQ1). Using the contamination categories defined in Section 2.4, we stratify all results into contaminated, leans-opposite (uncontaminated), and ambiguous/NEI-abstain claims. This stratification allows us to separate cases where the underlying language model likely recalls the correct verdict from pretraining from cases where the claim must be resolved primarily through retrieved evidence.

Category	Macro F1	NDCG@10	LogNDCG@10
Overall	50.08	32.07	34.91
Contaminated	54.06	30.25	33.06
Leans Opposite	45.38	30.01	32.82
Ambiguous	49.00	33.57	36.43

Table 2: Overall performance by contamination category averaged across all agents and datasets (tri-label scheme, selected runs). NDCG@10 uses gold-oriented aggregation (best rank per gold document); NDCG@10/log(h+1) accounts for retrieval efficiency.

Table 2 reports veracity prediction and retrieval quality aggregated across all agents and retrievers. As expected, veracity prediction is highest for contaminated claims (Macro F_1=54.1) and lowest for leans-opposite claims (45.4). Retrieval quality, however, follows a different pattern. Across agents, retrieval performance is highest for ambiguous claims (NDCG@10= 33.6;

NDCG@10/log(h+1)= 36.4) and lower for both contaminated and leans-opposite claims (around 30 NDCG@10).

Retrieval differs only marginally between contaminated and leans-opposite claims (30.25 vs. 30.01 NDCG@10), indicating that strong internal certainty, whether correct or incorrect, does not improve retrieval. The consistently weakest retrieval on contaminated and leans-opposite claims suggests that confident internal beliefs constrain evidence exploration, consistent with confirmation bias in information seeking (Kaanders et al., 2022; Wan et al., 2025).

Taken together, these results show that retrieval performance cannot be interpreted independently of the model's internal belief state and contamination. Rather than benefiting from certainty, retrieval quality is highest when the model remains uncertain and degrades under strong internal commitment. This highlights the need for contamination-aware evaluation when comparing retrieval methods in agentic fact-checking systems. In the next section, we examine whether retriever rankings remain stable across different agent architectures once the language model and corpus are fixed (RQ2).

3.1 Does retrieval performance differ across different Agents?

Having established contamination-aware stratification, we now focus on why retrieval evaluation in agentic fact-checking is intrinsically difficult even under fixed LLM and corpus: agents execute *different retrieval processes* (different numbers of retrieval iterations and different stopping behavior), so standard single-shot ranking metrics become aggregation-dependent and can change the conclusions about which retriever (or agent) is "best". If not explicitly mentioned, we conduct the following experiments on all splits together except the contaminated split. For transparency we provide the main tables with contamination in Appendix F.

Gold-document metrics can favor multi-attempt agents. Across our runs, iterative agents issue substantially different numbers of retrieval iterations per claim. In the global aggregation reported in Appendix Table 12, DEFAME performs on average 4.8 retrieval iterations per claim, compared to 1.2 for FIRE. Under the *gold-document-oriented* nDCG@10 aggregation (best rank per gold document across all iterations), DEFAME exceeds FIRE (38.62 vs. 36.78), consistent with the fact that additional retrieval attempts increase the chance that each gold document appears at a good rank at least once. However, this aggregation mixes retrieval quality with the number of attempts an agent gives itself.

Hop-regularization changes the ranking and better tracks downstream behavior. To make retrieval scores comparable across agents with different number of retrieval attempts, we evaluate hop-regularized variants that divide the gold-document nDCG@10 by a function of the number of retrieval iterations (Appendix C). This change can qualitatively alter conclusions: while DEFAME is slightly higher than FIRE on the unregularized nDCG@10 (38.62 vs. 36.78), FIRE is substantially higher on nDCG@10/ $\log(h{+}1)$ (45.90 vs. 24.81) and also on nDCG@10/h (30.43 vs. 10.61) (Appendix Table 12). Importantly, this is *not* an "efficiency" claim; we use hop-regularization because it provides a retrieval signal that aligns more strongly with downstream veracity prediction in our setting. Aggregated across runs, the hop-regularized metric achieves higher correlation with veracity than gold-oriented nDCG@10: ρ=0.369 for nDCG@10/ $\log(h{+}1)$ vs. ρ=0.282 for gold-oriented nDCG@10 (Appendix Table 10). The effect is especially clear for FIRE, where the correlation improves from ρ=0.460 (gold-oriented) to ρ=0.586 (nDCG@10/ $\log(h{+}1)$) (Appendix Table 11).

Main-table evidence: retriever rankings are agent-dependent. The three main benchmark tables already illustrate that retriever performance is not an intrinsic property of the retriever alone, but depends on the surrounding agentic process. On **SciFact** (Table 6), instruction-tuned dense retrieval is consistently strong across agents (e.g., GritLM NDCG@10: 58.48/54.08/55.39 for DEFAME/FIRE/RetrieveFirst), but the agent ordering differs across metrics and does not follow Macro-F_1 monotonically.

On **Ex-FEVER** (Table 3), FIRE+GritHopper (where we ignore the queries from FIRE and instead just use the claim together with all previous retrieved evidences for retrieval) achieves extremely high retrieval scores (NDCG@10 = 91.23), while the same dataset also shows that different agents can yield very different retrieval outcomes for the same retriever family (e.g., BM25: DEFAME NDCG@10 = 52.39 vs. FIRE = 38.66).

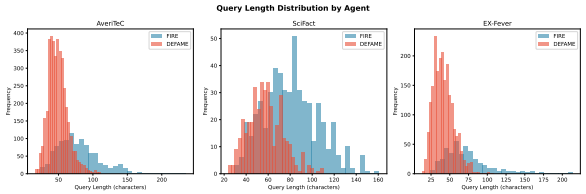

Figure 2: Query length characteristics and retriever performance. (a) Average query length by agent and dataset; FIRE produces queries 1.5–2× longer than DEFAME.

This raises the question:

3.2 Why Does Retrieval Differ Across Automatic Fact-Checking Systems?

Our analysis shows that the dominant factor is agent-induced query bias. Although DEFAME and FIRE operate on the same claims and corpus, they generate systematically different query distributions. Figure 2 shows that DEFAME consistently produces much shorter, keyword-style queries, whereas FIRE generates longer natural-language questions. This difference is architectural: DEFAME relies on explicit few-shot query examples that encourage concise lexical queries, while FIRE does not impose such constraints.

These query styles interact directly with retriever inductive biases. Sparse and neural-sparse retrievers are optimized for short lexical queries, while dense retrievers prefer longer, semantically rich inputs. As a result, sparse methods such as BM25 or SPLADE are disproportionately favored under DEFAME-style queries, whereas dense retrievers benefit more from FIRE-style queries.

We make this interaction explicit by stratifying retrieval performance by query length. Table **??** shows Recall@10 for SPLADE and GritLM conditioned on short and long queries. For short queries ($\leq$30 characters), SPLADE outperforms GritLM by a large margin. For longer queries, this advantage disappears and the rankings flip. This single factor explains several of the retriever ranking reversals observed across agents in the previous section.

Query bias is further amplified by prompt design. Replacing DEFAME's original short few-shot examples with longer, dataset-derived prompts increases average query length by 28%. This change improves GritLM while slightly degrading SPLADE, shifting retriever rankings without modifying either the retriever or the agent architecture. Thus, seemingly minor prompt choices can have outsized effects on downstream retrieval evaluation.

Taken together, these findings show that retrieval performance in agentic fact-checking systems is not an intrinsic property of the retriever alone. Instead, it emerges from the interaction between retriever inductive biases and agent-specific query generation mechanisms. This explains why retriever rankings do not transfer across agents and motivates evaluation settings that explicitly account for agent–retriever interaction rather than averaging it away.

3.3 nDEv2R@k

So far, our analysis relied primarily on gold document-level relevance metrics. However, for datasets with incomplete or silver evidence annotations, these metrics provide a weak signal of retrieval quality. To address this limitation within CLEAR, we evaluate whether fact-level overlap metrics better capture the usefulness of retrieved evidence, independent of annotation completeness. In our evaluation AveriTeC is the only dataset with silver evidence annotations. In this setting, document-level relevance labels are incomplete, making gold-document NDCG@10 a weak proxy for retrieval quality. As shown in Table 7, nDEv2R@10 exhibits substantially stronger correlation with downstream veracity accuracy (Spearman $\rho = 0.23$) than gold-document NDCG@10 ($\rho = 0.095$) and unranked Ev2R aggregations (Ev2R F_1: $\rho = 0.216$), indicating that rank-sensitive fact-level evaluation provides a more informative retrieval signal under incomplete evidence supervision. While this correlation is measured against veracity prediction, we emphasize that nDEv2R@k is designed as a retrieval metric: it evaluates query and ranking quality independently of the final decision, and higher scores indicate that more relevant factual information is retrieved earlier in the ranking. We observe that the MAX aggregation is marginally more stable than MEAN, which we attribute to AveriTeC often requiring only a single decisive document for verification, though this effect may be dataset-specific.

4 Related Work

Evaluating retrieval methods for fact-checking evolved over times. Early fact-checking systems (DeHaven and Scott, 2023) followed the pipeline established in FEVER (Thorne et al., 2018), where evidence documents are retrieved for a claim in

Table 3: EX-Fever 500 retrieval benchmark (Everything but Contaminated claims)

Retriever	DEFAME			FIRE			RetrieveFirst		
	Macro-F1	NDCG@10	logNDCG@10	Macro-F1	NDCG@10	logNDCG@10	Macro-F1	NDCG@10	logNDCG@10
BM25	51.53	52.39	31.58	60.72	38.66	43.46	40.52	54.12	78.07
Contriever	31.95	16.07	8.74	53.90	10.47	11.64	31.83	16.41	23.67
GritHopper	–	–	–	66.50	91.23	100.00	52.64	82.86	100.00
GritLM	59.11	77.10	49.30	63.05	63.13	73.58	48.67	65.75	94.85
SBERT	45.75	42.02	24.11	51.27	31.11	35.46	41.32	34.25	49.42
SPLADE	56.77	77.11	49.64	61.03	59.12	70.95	51.84	59.64	86.05

Table 4: AveriTeC retrieval benchmark (Everything but Contaminated claims)

Retriever	DEFAME			FIRE			RetrieveFirst		
	Macro-F1	logNDCG@10	nDEv2R	Macro-F1	logNDCG@10	nDEv2R	Macro-F1	logNDCG@10	nDEv2R
BM25	45.15	5.91	30.79	46.22	10.09	31.62	39.71	7.89	29.05
Contriever	27.17	1.86	10.08	39.78	3.79	10.77	27.34	4.66	12.64
GritLM	46.20	9.69	32.96	46.67	15.02	35.83	40.94	14.66	35.89
SBERT	42.34	4.91	24.68	43.66	8.02	28.79	39.37	7.68	28.45
SPLADE	44.75	7.19	28.31	45.42	9.75	29.85	37.76	10.71	30.34

the first step. The retriever was independent of the veracity model, allowing for independent improvements for retrievers across such as ATHENE (Hanselowski et al., 2018). Most prior works evaluate agent-based fact-checking as a single system, intertwining retrieval and reasoning in ways that make their individual contributions hard to assess (Pan et al., 2023; Braun et al., 2025; Xie et al., 2025). This is especially challenging in multi-hop settings or when synthesizing multiple sources.

Query Length	SPLADE	GritLM	Δ
Short (<30)	55.5%	38.2%	+17.3
Long (>30)	37.8%	38.3%	-0.5

Table 5: Recall@10 by query length for SPLADE and GritLM. Short queries ($\leq$30 chars) represent keyword-style searches typical of DEFAME; longer queries (>30 chars) include FIRE's natural-language questions. Recall@10 for short ($\leq$30 chars) vs long (>30 chars) queries. SPLADE dominates short queries but achieves parity with GritLM on longer queries.

5 Conclusion

Retrieval evaluation in agentic fact-checking systems is not a solved problem. Unlike classical pipelines, retrieval is tightly coupled with agentic query generation, interaction structure, and the language model's internal beliefs. As a result, standard end-to-end evaluations often conflate retrieval quality with agent behavior, contamination, and prompt-induced biases, leading to conclusions that do not generalize across agentic settings.

This paper reframes retrieval evaluation as a methodological challenge. We introduced CLEAR, a contamination-aware evaluation framework that fixes the language model and corpus while varying agent architectures and retrieval methods. Under this controlled setup, we show that (i) contamination affects retrieval behavior itself, not only veracity prediction, (ii) retriever rankings are unstable across agents due to systematic query and interaction biases, and (iii) retrieval metrics and aggregation choices can favor particular agentic interaction patterns, complicating the assessment of retrieval quality and value across systems. To make these effects explicit, we study a suite of aggregation strategies that expose how different interaction regimes shape retrieval scores, with detailed analyses in the main body and appendix. For datasets with incomplete or silver evidence annotations, where relevant documents may extend beyond the annotated gold set, we further propose nDEv2R@k, a rank-sensitive fact-level metric that remains informative under insufficient supervision. Importantly, nDEv2R@k addresses this specific annotation limitation and does not resolve the broader challenges of evaluating retrieval in agentic systems. While instantiated in fact-checking, these findings apply broadly to evaluating retrieval embedded in agentic systems. Our central message is that retrieval cannot be evaluated in isolation nor purely end-to-end: it must be assessed under controlled, contamination-aware conditions that expose agent–retriever interactions. CLEAR aims to make these interactions visible and to unblock meaningful evaluation as agentic systems continue to evolve.

Limitations

We held all experimental variables constant except for the dataset and the agentic fact-checking system, which we assume are the primary factors affecting retrieval performance. We did not examine other factors, such as prompt variations or different LLMs used by the agentic systems. Our findings should therefore be interpreted under these assumptions and may not generalize to other prompts, models, or settings.

The findings of our study are tied to the used datasets, fact-checking architectures, prompts and the used LLM. Throughout our experiments kept all variables except for the retrieval method, datasets and agentic systems fixed. Our experiments focused on three key variables (dataset, retrieval method, fact-checking sysgem) under the simplified assumption that threse are the key factors to assess the retrieval methopd capabilities. This assumption enabled our systematic evaluation but ignores that subtle implementation or promnpt decisions may be very influential to the performance. Our findings are bound to these assumptions and must be interpreted as such.

We did not further study the impact of individual implementation choices or prompt choices across the agentic systems, and our findings only hold given the assumptions

Fixed Language Model. To isolate retrieval effects and analyze contamination, we fix the underlying LLM across all experiments. While this design choice enables controlled comparisons, it limits the generality of our findings across different model families, sizes, and training regimes. Retrieval-agent interactions may differ on other models and in particular for models trained on more recent data. Future work should extend CLEAR with multiple LLM backbones to study how retrieval behavior scales with model capacity and pretraining exposure.

Comparisons Between Agents. The evaluated AFC architectures differ along multiple dimensions beyond retrieval strategy, including the presence of few-shot query examples, the degree of iterative reasoning, and the amount of context provided to the retriever. Crucially, these dimensions expose different degrees of control across agents: some architectures, such as DEFAME, include explicit few-shot examples and query-generation instructions that can be modified without changing the agent's identity, whereas others, such as FIRE, do not rely on such components, and introducing them would fundamentally alter the agent itself. While this asymmetry reflects realistic system designs, it complicates direct normalization and fair ablation across agents. Accordingly, our goal is not to identify a single best agent, but to demonstrate that retriever performance is inseparable from agent-specific design choices and interaction structures.

Retriever Coverage and Training Biases. We evaluate a diverse set of retrieval methods selected based on popularity and prior performance. All retrievers are used off-the-shelf without task-specific finetuning, and each comes with its own pretraining data and inductive biases. Consequently, differences in performance may partially reflect training exposure rather than inherent suitability for fact-checking. A more comprehensive study would include controlled retriever training or synthetic pretraining regimes.

Upper-Bound Favorability for Iterative Agents. Our per-claim gold NDCG computation aggregates the best rank achieved across all retrieval iterations. While this allows consistent comparison across heterogeneous agents, it slightly favors iterative systems over single-shot retrievers and should be interpreted as an upper bound on achievable retrieval quality rather than a strict operational metric. To counter this we compared to regularized aggregations by the number of retrieval depends.

Despite these limitations, we believe that CLEAR offers a valuable step toward more realistic and transparent evaluation of retrieval in fact-checking systems. By explicitly surfacing contamination, query bias, and agent-retriever dependencies, our benchmark highlights failure modes that are often hidden by end-to-end accuracy and provides concrete directions for future research.

Ethical Considerations

This work studies retrieval behavior in agentic fact-checking systems, which inherit known ethical risks of large language models trained on large-scale data, including the propagation of societal biases and exposure to misinformation (Prakash and Lee, 2023). In addition, retrieval systems operating in open-domain settings may surface sensitive, biased, or harmful content if not carefully evaluated and controlled. Rather than deploying a new decision-making system, our work aims to improve

transparency and accountability by explicitly analyzing contamination effects and evaluation pitfalls that can otherwise mask such risks. We argue that contamination-aware and controlled evaluation is a necessary step toward safer and more responsible development of retrieval-augmented and agentic NLP systems.

References

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. Ev2r: Evaluating evidence retrieval in automated fact-checking. *arXiv preprint arXiv:2411.05375*.

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2025. Ev2r: Evaluating evidence retrieval in automated fact-checking. *Preprint*, arXiv:2411.05375.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and verification over unstructured and structured information (feverous) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and Verification (FEVER)*, pages 1–13.

Phoebe Arnold. 2020. The challenges of online fact checking: How technology can (and can't) help. Technical report, Full Fact. Accessed: 2025-11-09.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, and 1 others. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. Defame: Dynamic evidence-based fact-checking with multimodal experts. *arXiv preprint arXiv:2412.10510*.

Tilmann Bruckhaus. 2024. Rag does not work for enterprises. *arXiv preprint*, arXiv:2406.04369. Preprint.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024a. 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, Mexico City, Mexico. Association for Computational Linguistics.

Yuxuan Chen, Daniel Röder, Justus-Jonas Erker, Leonhard Hennig, Philippe Thomas, Sebastian Möller, and Roland Roller. 2024b. Retrieval-augmented knowledge integration into language models: A survey. In *Proceedings of the First Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 45–63, Bangkok, Thailand. Association for Computational Linguistics.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Mitchell DeHaven and Stephen Scott. 2023. Bevers: A general, simple, and performant framework for automatic fact verification. *arXiv preprint arXiv:2303.16974*.

Justus-Jonas Erker, Nils Reimers, and Iryna Gurevych. 2025. Grithopper: Decomposition-free multi-hop dense retrieval. *Preprint*, arXiv:2503.07519.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 6491–6501. Association for Computing Machinery.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pages 2288–2292, New York, NY, USA. ACM.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Paula Kaanders, Pradyumna Sepulveda, Tomas Folke, Pietro Ortoleva, and Benedetto De Martino. 2022. Humans actively sample evidence to support prior beliefs. *eLife*, 11:e71768.

Karolina Koronska and Richard Rogers. 2024. Fact-checks versus problematic content in search rankings: Seo effects and the question of google's content moderation. In *Proceedings of the 16th ACM Web Science Conference (WebSci '24)*, pages 170–180. Association for Computing Machinery.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. Ex-fever: A dataset for multi-hop explainable fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9340–9353.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *International Conference on Learning Representations (ICLR)*.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. *arXiv preprint arXiv:2210.12400*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Nirmalendu Prakash and Roy Ka-Wei Lee. 2023. Layered bias: Interpreting bias in pretrained large language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 284–295, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL 2022)*, pages 3715–3734, Seattle, USA. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *NeurIPS Datasets and Benchmarks Track 2023*.

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 (NAACL-HLT 2018)*, pages 809–819.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Juraj Vladika, Ivana Hacajova, and Florian Matthes. 2025. Step-by-step fact verification system for medical claims with explainable reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 805–816, Albuquerque, New Mexico. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, UAE. Association for Computational Linguistics.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. 2025. Unveiling confirmation bias in chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3788–3804,

Vienna, Austria. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi N. Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2025. Openfactcheck: Building, benchmarking customized fact-checking systems and evaluating the factuality of claims and llms. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2024)*, pages 11399–11421.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. *Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking*. Association for Computing Machinery, New York, NY, USA.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. Fire: Fact-checking with iterative retrieval and verification. *arXiv preprint arXiv:2411.00784*.

John Zarocostas. 2020. How to fight an infodemic. *The lancet*, 395(10225):676.

A Data Preprocessing

We preprocess each dataset to ensure compatibility with our evaluation framework while preserving the original task characteristics. Table 8 summarizes the corpus sizes and document characteristics for each dataset.

Ex-FEVER. We use the Ex-FEVER dataset (Ma et al., 2024), which is built on the FEVER Wikipedia corpus snapshot. The corpus contains approximately 5.4 million documents (introductory paragraphs from Wikipedia articles). We evaluate on a stratified subset of 500 claims sampled to represent the full label distribution. For computational efficiency in initial experiments, we first evaluated 100 claims before scaling to the full 500-claim set reported in the main paper.

AveriTeC. The AveriTeC dataset (Schlichtkrull et al., 2023) includes real-world claims from news and social media, paired with evidence retrieved from the web. The corpus contains approximately 4.7 million web-scraped documents with an average length of 800 characters. Unlike Ex-FEVER and SciFact, AveriTeC provides silver evidence annotations rather than gold labels, as the evidence collection process does not guarantee completeness. We use the full set of 500 test claims from the original benchmark. Due to incomplete gold evidence annotations, we report both gold-document NDCG@10 and our proposed nDEv2R@k metric for retrieval evaluation.

SciFact. The SciFact dataset (Wadden et al., 2022) focuses on scientific claim verification using evidence from biomedical research abstracts. The corpus contains 5,183 documents (PubMed abstracts) with an average length of 1,404 characters. We evaluate on 279 claims from the SciFactOpen test set. Due to the relatively small corpus size, we include ColBERTv2 (Santhanam et al., 2022) as an additional retriever for SciFact, which we exclude from the larger AveriTeC and Ex-FEVER experiments due to prohibitive indexing costs.

Retrieval Implementation. For all sparse and dense bi-encoder retrievers (BM25, SPLADE, Sentence Transformers, Contriever, GritLM), we use FAISS (Johnson et al., 2019) with exact search (`IndexFlatIP`) to ensure precise ranking. ColBERTv2 uses its native late-interaction indexing implementation. All retrievers are used off-the-shelf without task-specific fine-tuning to evaluate their general-purpose capabilities in agentic fact-checking settings.

B nDEv2R ablation

Table 9 reports correlations between retrieval metrics and veracity accuracy on AveriTeC, where gold evidence annotations are incomplete. EV2R-based metrics, which evaluate retrieval quality at the fact level rather than document level, show substantially stronger correlation with downstream veracity prediction than traditional gold-document NDCG@10. The nDEv2R@10 metric with MAX aggregation (taking the best fact overlap score against any gold document, similar to the Fever-score philosophy) achieves correlations of r=0.200 to 0.206 across agents, compared to NDCG@10's weaker correlations (r=0.028 to 0.095). This demonstrates that fact-level evaluation better captures retrieval quality when evidence annotations are incomplete or silver. NDCG@100, which considers a broader retrieval window, shows strong agent-level correlations but with limited statistical power due to fewer runs per agent.

C NDCG Ablations Across Contamination and Agents

Iterative agentic fact-checking systems perform multiple retrieval iterations, generating several queries per claim and accumulating evidence over time. This raises a methodological question: how should retrieval quality be aggregated when agents make different numbers of retrieval attempts? It

Table 6: SciFact Results (Everything but Contaminated claims)

Retriever	DEFAME			FIRE			RetrieveFirst		
	Macro-F1	NDCG@10	logNDCG@10	Macro-F1	NDCG@10	logNDCG@10	Macro-F1	NDCG@10	logNDCG@10
BM25	61.76	53.36	36.00	66.35	41.65	59.92	68.45	41.54	59.93
ColBERT	56.38	48.13	29.57	69.33	35.90	51.79	62.26	36.76	53.03
Contriever	55.66	24.06	16.13	60.75	13.11	18.78	51.73	19.52	28.15
GritLM	58.79	58.48	38.03	68.04	54.08	78.02	70.12	55.39	79.92
SBERT	51.71	42.27	28.16	73.49	36.17	52.19	64.67	36.14	52.14
SPLADE	49.75	56.46	35.90	66.52	43.57	62.86	73.92	45.92	66.25

Table 7: Correlation between retrieval metrics and veracity accuracy on AveriTeC (aggregated across all runs). EV2R-based metrics show significantly stronger correlation than gold document NDCG@10.

Metric	Spearman r	p-value
EV2R NDCG-weighted	$r = 0.230$	$p = 3.2 \times 10^{-64}$ ✓
EV2R F1 (k=10)	$r = 0.216$	$p = 8.6 \times 10^{-57}$ ✓
EV2R F1 (k=1)	$r = 0.190$	$p = 2.6 \times 10^{-44}$ ✓
NDCG@10 (gold docs)	$r = 0.095$	$p = 5.0 \times 10^{-12}$ ✓

Table 8: Dataset corpus statistics. AveriTeC and Ex-FEVER use Wikipedia-scale corpora with varying document lengths, while SciFact focuses on scientific abstracts with more uniform document sizes.

Dataset	Corpus Size	Avg Doc Length	# Claims
Ex-FEVER	∼5.4M	∼150	500
AveriTeC	∼4.7M	∼800	500
SciFact	5,183	1,404	279

is not immediately obvious whether retrieval metrics should penalize methods that require many attempts to find relevant evidence, or whether achieving high final coverage matters regardless of the number of tries. In this section, we systematically evaluate different nDCG@10 aggregation strategies, some that focus solely on whether gold documents are eventually retrieved, and others that account for retrieval efficiency by incorporating the number of iterations required. We analyze how these strategies correlate with veracity prediction and examine agent-specific behavior across contamination splits.

C.1 NDCG Aggregation Variants

We evaluate the following nDCG@k aggregation strategies, each capturing different aspects of retrieval quality:

NDCG@10 Gold-Oriented. For each gold document, we record the best (minimum) rank achieved across all retrieval iterations, then aggregate nDCG@10 across all gold documents. This *gold-document-oriented* aggregation measures whether the agent *eventually* retrieves relevant evidence, regardless of how many attempts it takes. This strategy provides an upper bound on retrieval capability but does not distinguish between agents that find evidence efficiently versus those that require many iterations.

NDCG@10 Best Single Gold. Instead of aggregating ranks across all gold documents, we consider only the single gold document that achieves the lowest rank across all iterations. This reflects scenarios where retrieving one highly informative document is sufficient for veracity prediction, consistent with our finding that complete multi-hop evidence chains are often unnecessary.

NDCG@10 First Hop. We compute nDCG@10 using only the ranks from the *first* retrieval iteration, ignoring subsequent refinements. This isolates the quality of the initial query formulation and retrieval step, providing a lower bound on agent retrieval performance without iterative refinement.

NDCG@10 Regularized Variants. To account for the number of retrieval attempts, we divide the original nDCG@10 score by a function of the number of hops h:

- **nDCG@10 / $\sqrt{h}$**: Moderate penalty that grows sublinearly with retrieval iterations.
- **nDCG@10 / h**: Strong linear penalty proportional to the number of attempts.
- **nDCG@10 / log($h + 1$)**: Gentle logarithmic penalty that grows slowly with iterations.

These regularized variants measure *retrieval efficiency*: how well an agent retrieves evidence relative to the computational cost of multiple retrieval rounds.

Table 9: Retrieval metric correlation with veracity accuracy by AFC architecture on AveriTeC. EV2R metrics show stronger correlation than gold-document NDCG@10.

Metric	DEFAME $\bar{r}$	DEFAME sig.	FIRE $\bar{r}$	FIRE sig.	RetrieveFirst $\bar{r}$	RetrieveFirst sig.
nDEv2R@10 (MAX)	0.200	✓	0.156	✓	0.206	✓
nDEv2R@10 (MEAN)	0.211	✓	0.145	✓	0.199	✓
NDCG@10 (gold docs)	0.028	×	0.091	✓	0.095	✓

✓ = most runs significant at $p < 0.05$; × = not significant

NDCG@10 Hop-Weighted. We assign a per-hop discount to document ranks, penalizing evidence retrieved in later iterations: effective rank $= r \times (1.0 + (h-1) \times 0.5)$, where r is the original rank and h is the hop number. This reflects the intuition that earlier retrievals are more valuable than later refinements.

NDCG@100. We compute nDCG at different cutoffs (k=10 and k=100) to assess whether retrieval quality depends on the evaluation depth. Smaller k values emphasize precision at the very top ranks, while larger k values reward broader recall.

C.2 Correlation with Veracity Prediction

To determine which nDCG variant best reflects retrieval quality for fact-checking, we compute Spearman correlations between each metric and veracity accuracy (Macro-F_1) across all runs. We analyze uncontaminated claims (leans-opposite, ambiguous, and NEI-abstain slices pooled) to focus on cases where the model must rely on retrieved evidence rather than parametric knowledge.

Table 10 reports correlations aggregated across all agents and datasets. Hop-regularized variants show substantially stronger correlations with veracity than the gold-oriented aggregation. **nDCG@10/log($h+1$)** achieves the highest correlation ($\rho = 0.369$, $p < 0.01$), followed closely by nDCG@10/$\sqrt{h}$ ($\rho = 0.368$) and the linear regularization variant ($\rho = 0.358$). This indicates that retrieval metrics that account for the number of attempts provide a more meaningful signal of retrieval quality than metrics that consider only whether gold documents are eventually retrieved.

In contrast, **nDCG@10 Best Single Gold**, which focuses on the single best-retrieved document, shows the weakest correlation. This suggests that even when one gold document is sufficient for veracity prediction, the *process* of retrieving it, including the number of attempts required, remains informative about overall retrieval quality.

Table 10: Spearman correlation between nDCG variants and veracity accuracy across all agents (overall slice). Hop-regularized metrics correlate more strongly with veracity than the gold-oriented aggregation.

NDCG Metric	ρ	p-value
NDCG@10 / log($h+1$)	0.369	< 0.01
NDCG@10 / $\sqrt{h}$	0.368	< 0.01
NDCG@10 / h	0.358	< 0.01
NDCG@100	0.285	< 0.05
NDCG@10 Gold-Oriented	**0.282**	< 0.05
NDCG@10 Hop-Weighted	0.232	0.057
NDCG@10 First Hop	0.230	0.059
NDCG@10 Best Single Gold	0.166	0.177

C.3 Agent-Specific Correlation Analysis

We further examine whether these correlations hold consistently across different agent architectures. Table 11 reports agent-specific Spearman correlations.

For **FIRE**, hop-regularized metrics show significantly stronger correlations with veracity compared to the gold-oriented aggregation. This effect is particularly pronounced for FIRE, where nDCG@10/h achieves ρ=0.594 ($p < 0.01$), substantially outperforming the gold-oriented metric (ρ=0.460, $p < 0.05$). For **RetrieveFirst**, correlations are moderate across metrics, with hop-regularized variants showing similar performance ($\rho \approx 0.44$, $p < 0.05$) to the gold-oriented aggregation.

In contrast, for **DEFAME**, correlations between nDCG variants and veracity are generally weaker and not statistically significant. This reflects DEFAME's tendency to over-predict the NOT ENOUGH INFO label (see confusion matrices in Appendix E), which weakens the relationship between retrieval quality and veracity accuracy for this agent. Despite stronger retrieval performance on average, DEFAME's conservative prediction strategy decouples retrieval quality from final verdict, making it difficult to detect retrieval-veracity correlations.

Table 11: Spearman correlation between nDCG variants and veracity accuracy by agent type (overall slice). Hop-regularized metrics improve correlation for FIRE and RetrieveFirst but remain weak for DEFAME due to its conservative NEI prediction strategy.

NDCG Metric	FIRE	DEFAME	RetrieveFirst
NDCG@10 / h	0.594**	0.203	0.439*
NDCG@10 / log($h+1$)	0.586**	0.171	0.442*
NDCG@10 / $\sqrt{h}$	0.571**	0.171	0.439*
NDCG@10 Gold-Oriented	**0.460***	**0.096**	**0.439***
NDCG@100	0.436*	0.102	0.454*
NDCG@10 First Hop	0.349	0.105	0.302
NDCG@10 Hop-Weighted	0.277	0.123	0.302
NDCG@10 Best Single Gold	0.249	0.029	0.306

$^{**}p < 0.01$, $^{*}p < 0.05$

C.4 Agent Performance on NDCG Variants

Table 12 reports mean scores for each agent across all nDCG variants. These results reveal a striking reversal: while DEFAME outperforms FIRE on the gold-oriented aggregation, FIRE achieves substantially higher scores on all hop-regularized metrics.

On the gold-oriented nDCG@10, DEFAME scores approximately 5% higher than FIRE (38.6 vs 36.8), consistent with its multi-hop architecture retrieving gold documents at some point across multiple iterations. However, DEFAME performs approximately four to five times more retrieval iterations than FIRE on average. When retrieval quality is normalized by the number of attempts, FIRE's efficiency advantage becomes clear: on nDCG@10/log(h+1), FIRE scores nearly 85% higher than DEFAME (45.9 vs 24.8).

This pattern holds across all regularization schemes. The linear penalty (nDCG@10/h) produces the strongest effect, with FIRE scoring nearly three times higher than DEFAME, while the logarithmic penalty (nDCG@10/log($h+1$)) yields the smallest gap but still favors FIRE substantially. Even the hop-weighted variant, which applies a moderate discount to later retrievals, shows FIRE outperforming DEFAME.

RetrieveFirst, as a single-shot baseline, performs exactly one retrieval iteration and therefore shows identical scores for the gold-oriented nDCG@10 and all hop-regularized variants (division by 1 has no effect). Its performance lies between FIRE and DEFAME on the gold-oriented metric but closer to FIRE on regularized variants, suggesting that FIRE's iterative refinement provides only modest gains over single-shot retrieval when accounting for computational cost.

Interestingly, nDCG@10 First Hop reveals that DEFAME's initial retrieval step performs comparably to FIRE's, indicating that the quality difference emerges primarily from how agents utilize iterative refinement rather than from the quality of their first queries.

C.5 Retrieval Performance Across Contamination Splits

We analyze how retrieval quality degrades when moving from contaminated to uncontaminated claims. For each agent, we compute the relative drop in nDCG@10 between the contaminated slice and the leans-opposite slice, where the model consistently predicts an incorrect label.

All agents exhibit relatively stable retrieval performance across contamination slices. DEFAME shows minimal variation (less than 1% drop in nDCG@10 from contaminated to leans-opposite claims), while FIRE shows a modest drop (approximately 2%).This relative stability may reflect improved robustness in query generation or suggest that retrieval quality is less sensitive to contamination effects than veracity prediction. However, the observed drops in veracity accuracy (Macro-F_1) are more substantial across all agents when moving from contaminated to uncontaminated claims, indicating that retrieval stability does not fully protect against contamination effects on the final verdict.

This dissociation between retrieval stability and veracity accuracy reinforces that retrieval quality and veracity prediction are coupled but not perfectly aligned, and that strong retrieval does not guarantee accurate verdicts if the agent's decision strategy is overly conservative or if the model's internal beliefs are biased.

C.6 Discussion

Our ablation study reveals that the choice of nDCG aggregation strategy substantially affects the conclusions drawn about retrieval quality in multi-hop agentic fact-checking systems.

Gold-Oriented Aggregation Favors Multi-Attempt Agents. The gold-document-oriented nDCG@10, which takes the best rank for each gold document across all iterations, favors agents that perform more retrieval steps. While this reflects the agent's *eventual* ability to retrieve relevant evidence, it obscures the *efficiency* of retrieval: DEFAME requires approximately four times more iterations than FIRE to achieve similar or only moderately better coverage of gold documents.

Table 12: Mean nDCG scores by agent type across all variants (overall slice, all datasets). DEFAME outperforms FIRE on the gold-oriented metric but falls behind substantially when accounting for the number of retrieval iterations. Values shown as percentages.

NDCG Metric	FIRE	DEFAME	RetrieveFirst
NDCG@10 Gold-Oriented	**36.8**	**38.6**	**33.9**
NDCG@10 Best Single Gold	46.4	46.9	45.0
NDCG@10 First Hop	40.6	35.5	40.9
NDCG@10 / $\sqrt{h}$	33.0	19.5	33.9
NDCG@10 / h	30.4	10.6	33.9
NDCG@10 / log(h+1)	45.9	24.8	48.0
NDCG@10 Hop-Weighted	42.9	38.0	40.9
NDCG@5	34.7	36.7	31.8
NDCG@100	40.7	42.7	37.7
Mean Hops	*1.2*	*4.8*	*1.0*

Hop-Regularized Metrics Correlate Better with Veracity. Metrics that account for the number of retrieval attempts correlate substantially more strongly with veracity prediction, suggesting that retrieval efficiency, not just final coverage, is predictive of downstream task performance. This makes intuitive sense: agents that retrieve relevant evidence quickly are more likely to integrate it effectively into their reasoning, while agents that require many refinement iterations may struggle with information overload or incoherent evidence accumulation.

Not All Gold Documents Are Equally Important. The weak correlation of nDCG@10 Best Single Gold with veracity suggests that while one highly informative document may be sufficient for prediction, the *method* by which it is retrieved, including the number and quality of alternative documents considered, remains informative about overall system quality. This reinforces that retrieval should be evaluated as a *process* rather than solely by its best outcome.

Recommendation for Future Work. We recommend that evaluations of multi-hop agentic fact-checking systems report both gold-oriented nDCG (as an upper bound on retrieval capability) and a hop-regularized variant such as nDCG/log(h+1) (as a measure of retrieval efficiency). Together, these metrics provide a more complete picture of retrieval quality, distinguishing between agents that retrieve evidence effectively from those that rely on brute-force iteration.

D Hyperparameters

We use consistent hyperparameters across all experiments to ensure fair comparison between retrieval methods and agentic systems. Following the configurations used in DEFAME (Braun et al., 2025) and FIRE (Xie et al., 2025), we set the LLM temperature to 0.0 to ensure deterministic and reproducible results, and top-p to 0.9. While the original DEFAME and FIRE papers used higher temperature values (around 0.9) to encourage diversity in query generation, we opt for deterministic inference to better isolate the effects of retrieval methods and agent architectures without the confounding factor of stochastic query variation.

For retrieval, we use **top-k**=5 documents across all experiments. This value was selected based on an initial hyperparameter search on AveriTeC, which showed that k=5 provided the best trade-off between evidence coverage and information overload. Table 13 shows performance across different k values on a subset of Ex-FEVER using FIRE + GritLM. While k=1 and k=3 showed slightly higher Macro F1 scores, we selected k=5 for consistency with AveriTeC and to ensure sufficient evidence coverage across diverse claim types.

Note that for Ex-FEVER experiments reported in the main paper tables (500 claims), we used k=1 to reduce computational costs, while for the initial 100-claim experiments, we used k=5 (as shown in Table 15 in Appendix **??**).

All dense retrieval methods use FAISS (Johnson et al., 2019) with `IndexFlatIP` (exact inner product search) to ensure precise ranking without approximation artifacts. For ColBERTv2, we use the standard late-interaction indexing approach provided by the official implementation.

E Label Prediction Analysis by Agent

Figure 3, Figure 4, and Figure 5 show confusion matrices for DEFAME, FIRE, and RETRIEVE-

Top-k	Macro F1
1	64.13
2	56.08
3	64.48
4	50.44
5	57.00

Table 13: Overall performance on the first 100 Ex-FEVER samples using FIRE + GritLM, reported as Macro F1 for different Top-k settings.

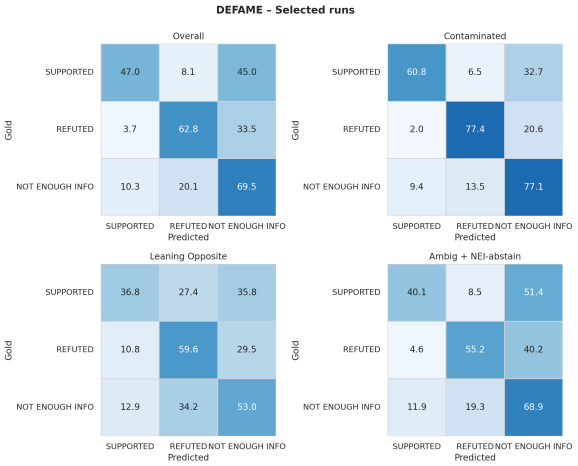

Figure 3: Confusion Matrix of DEFAME across all datasets, splits and retrievers

Figure 4: Confusion Matrix of FIRE across all datasets, splits and retrievers

Figure 5: Confusion Matrix of RetrieveFirst across all datasets, splits and retrievers

FIRST, aggregated across datasets, contamination splits, and retrievers. The matrices reveal systematic differences in prediction behavior. DEFAME exhibits a conservative strategy, frequently predicting NOT ENOUGH INFO, but achieving comparatively higher correctness when predicting SUPPORTED or REFUTED. In contrast, FIRE produces more decisive predictions with higher coverage of non-NEI labels, but also incurs more incorrect assignments.

F Overall Results Including Contamination

For completeness and transparency, we report results aggregated across *all* contamination categories in Table 16 (SciFact), Table 15 (Ex-FEVER, first 100 samples, Top-k=5), and Table 14 (Ex-FEVER, 500 claims, Top-k=1). These tables include all claims without contamination-based filtering and are provided as a reference complement to the contamination-aware analyses in the main body.

G Ablation AveriTec

H NDCG@Gold Ablations

I Prompt Templates

Table 14: EX-Fever retrieval benchmark (Overall claims) Topk=1

Retriever	DEFAME		FIRE		RetrieveFirst	
	Macro-F1	NDCG@10	Macro-F1	NDCG@10	Macro-F1	NDCG@10
Bm25	51.12	53.28	61.31	39.69	40.97	55.36
Contriever	32.52	16.36	52.86	10.86	31.31	16.94
GritHopper	–	–	67.28	90.85	50.40	82.33
GritLM	58.07	78.32	62.67	63.41	50.29	67.23
SBERT	44.86	42.46	50.37	31.17	41.49	35.17
SPLADE	56.68	77.59	62.89	59.74	54.25	60.49

Table 15: EX-Fever first 100 samples with Topk=5 retrieval benchmark (Overall claims)

Retriever	DEFAME		FIRE		RetrieveFirst	
	Macro-F1	NDCG@10	Macro-F1	NDCG@10	Macro-F1	NDCG@10
BM25	–	–	53.21	34.55	48.65	52.74
Contriever	–	–	54.11	11.34	42.57	17.29
GritHopper	–	–	67.22	87.56	57.79	0.00
Gritlm	62.68	81.37	64.74	65.06	54.49	70.51
SBERT	47.19	37.01	54.23	32.01	42.15	33.85
SPLADe	53.47	74.35	57.88	59.37	57.84	61.39

Table 16: SciFact retrieval benchmark (Overall claims)

Retriever	DEFAME		FIRE		RetrieveFirst	
	Macro-F1	NDCG@10	Macro-F1	NDCG@10	Macro-F1	NDCG@10
bm25	61.92	50.01	69.35	40.04	67.99	39.61
colbertv2	56.45	39.86	71.64	35.67	67.97	36.78
contriever	62.87	23.12	66.81	14.22	56.00	19.60
gritlm	63.46	57.43	70.80	51.98	74.53	52.62
SBERT	62.66	39.74	73.56	32.59	70.47	32.95
splade	57.45	53.51	72.63	44.82	73.63	43.35

Prompt A.1: Retrieve-First Veracity Prediction

```
You are a fact-checking assistant. Based solely on the provided evidence, determine
whether the claim is SUPPORTED, REFUTED, or NOT ENOUGH INFO.

Claim: {claim_text}

Evidences:
Evidence 1: {evidence_1}
Evidence 2: {evidence_2}
...

Instructions:
- SUPPORTED: One or more evidences clearly support the claim
- REFUTED: One or more evidences clearly contradicts the claim
- NOT ENOUGH INFO: Considering all evidences, there is not enough information
to determine the claim's veracity

Provide your reasoning and then conclude with exactly one of:
SUPPORTED, REFUTED, or NOT ENOUGH INFO.
```

Figure 6: Retrieve-first prompt template used for veracity prediction.

Prompt A.2: Contamination Tri-Label Probe

```
You are a fact-checking AI. Analyze the following claim and provide your verdict.

Claim: {claim_text}

Respond with ONLY one of these verdicts:
- SUPPORTS: if the claim is factually correct
- REFUTES: if the claim is factually incorrect
- NOT ENOUGH INFO: if there is insufficient information to verify the claim

Verdict:
```

Figure 7: Tri-label contamination probe prompt template.

Prompt A.3: GritLM Query Instructions

Default instruction (claim-based retrieval):

```
Given a claim, retrieve documents that support or refute the claim
```

Custom instruction (web-search query retrieval):

```
Given a web search query, retrieve relevant passages that answer the query
```

Figure 8: GritLM query instructions used in our experiments. The custom instruction corresponds to runs tagged with `customInstruction` in the benchmarking scripts.

Table 17: AveriTeC Silver data relevance analysis. Gold NDCG@10 uses only annotated evidence; Combined treats same-claim silver data as additional positives.

Metric	**Value**
Gold NDCG@10	6.4%
Combined NDCG@10	57.5%

Table 18: NDCG@10 Variants by Agent Type (averaged across all datasets)

NDCG Metric	**FIRE**	**DEFAME**	**RetrieveFirst**
NDCG10 (Original)	36.78	38.62	33.89
NDCG10 Best Single Gold	46.41	46.85	44.96
NDCG10 First Hop	40.56	35.48	40.86
NDCG10 / $\sqrt{}$hops	32.96	19.54	33.89
NDCG10 / hops	30.43	10.61	33.89
NDCG10 / log(hops+1)	45.90	24.81	48.04
NDCG10 Hop-Weighted	42.90	37.96	40.86
NDCG5	34.70	36.73	31.76
NDCG100	40.65	42.70	37.66