GQC: LLM-Based Grouped QA Consolidation for Open-Domain Fact Verification at AVeriTeC

Dongzhuoran Zhou[1,2],**Roxana Pop**[2], **Yuqicheng Zhu**[1,3], **Evgeny Kharlamov**[1,2],
[1]Bosch Center for AI
[2]University of Oslo
[3]University of Stuttgart
dongzhuoran.zhou@de.bosch.com, roxanap@ifi.uio.no,
yuqicheng.zhu@de.bosch.com, evgeny.kharlamov@de.bosch.com

Abstract

Structured fact verification benchmarks like AVeriTeC decompose claims into QA pairs to support fine-grained reasoning. However, current systems generate QA pairs independently for each evidence sentence, leading to redundancy, drift, and noise. We introduce a modular LLM-based QA consolidation module that jointly filters, clusters, and rewrites QA pairs at the claim level. Experiments show that this method improves evidence quality and veracity prediction accuracy. Our analysis also highlights the impact of model scale and alignment on downstream performance.

1 Introduction

Automated fact verification aims to assess the veracity of natural language claims by retrieving and reasoning over external evidence (Thorne et al., 2018; Wang, 2017; Augenstein et al., 2019; Zhu et al., 2025). While early systems typically treat this as a binary or multi-class classification problem using retrieved evidence as input, recent benchmarks—notably AVeriTeC (Schlichtkrull et al., 2023)—have introduced a structured pipeline where systems first generate clarification question–answer (QA) pairs based on retrieved evidence, and then use these QA pairs as an intermediate reasoning scaffold for final veracity prediction.

This structured QA paradigm improves transparency and evaluation granularity, but also introduces new challenges. While structured QA pipelines enable interpretability, they introduce new challenges. Systems like HerO (Yoon et al., 2024) generate QA pairs independently per sentence, resulting in overlapping or off-topic content that may confuse the final verifier. This redundancy inflates input length and can suppress relevant evidence. We propose a claim-level consolidation module to address these limitations and improve precision without sacrificing recall.

To address these issues, we introduce a simple and modular post-processing module that filters, clusters, and rewrites QA pairs using a large language model (LLM). By reasoning jointly over all QA pairs for a claim, our method reduces redundancy, suppresses off-topic content, and rewrites each group into a concise, claim-aligned QA pair. Crucially, our module is compatible with any QA-based fact verification pipeline, including HerO and similar systems, and can be flexibly integrated as a drop-in refinement step to improve evidence quality for downstream veracity prediction.

Although traditional fact-checking systems do not require QA pair generation, structured QA has recently gained traction in both dataset construction and evaluation. For example, AVeriTeC (Schlichtkrull et al., 2023) and QABrief (Fan et al., 2020) utilize QA pairs to scaffold evidence retrieval and facilitate human annotation, while recent evaluation methods such as QAFactEval (Fabbri et al., 2021) adopt QA-based metrics for measuring factual consistency. Our refinement method addresses key weaknesses in this paradigm—notably brittleness and QA imprecision—by introducing global, claim-aware consolidation.

Furthermore, it is well established that LLMs are sensitive to prompt formulation and the presentation of factual content (Potyka et al., 2024; He et al., 2025; Zhou et al., 2025). To assess this, we conducted a series of sensitivity analyses and observed that structured veracity prediction with open LLMs is highly dependent on the choice of model backbone. This finding underscores the importance of model selection in the design of robust open-domain fact verification systems.

The main contributions of this paper are as follows. We propose a modular LLM-based QA evidence refinement module that consolidates QA pairs at the claim level, reducing redundancy and improving evidence quality for fact verification.

Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), pages 151–161

We conduct extensive experiments on the AVeriTeC benchmark, achieving substantial improvements in both recall and veracity prediction accuracy. Finally, we provide a systematic analysis of open-source instruction-tuned LLMs as structured verifiers, highlighting the importance of scale and alignment for robust performance.

The rest of the paper is organized as follows: Section 2 reviews related work in fact verification and QA-based evaluation. Section 3 details our proposed QA evidence refinement methodology. Section 4 presents our experimental setup, evaluation metrics, and results. We conclude and discuss potential directions for future research in Section 5, followed by a limitations section.

2 Related Work

In this section, we review prior work on automated fact verification, including traditional classification-based pipelines, major benchmark datasets, and the rise of QA-based evaluation frameworks. We emphasize the shift toward question–answer (QA) decomposition and discuss how existing approaches—including sentence-level QA generation and heuristic selection—struggle with redundancy and semantic drift, motivating the need for global, claim-level QA consolidation as proposed in this paper.

Fact Verification Pipelines. Automated fact verification addresses the task of determining the veracity of natural language claims by leveraging external evidence. Early systems (Thorne et al., 2018; Augenstein et al., 2019; Wang, 2017) cast this as a classification problem: given a claim and retrieved evidence, the system predicts a veracity label. Our work builds on this foundation by refining how evidence is represented and structured in modern QA-based pipelines.

Benchmarks. Benchmarks. Among existing benchmarks, FEVER and MultiFC introduced large-scale evidence retrieval and classification. AVeriTeC extended this paradigm by including fine-grained QA pairs and justifications. Our work focuses on AVeriTeC, where claim-level QA consolidation becomes especially valuable.

QA-based Fact Verification. Structuring fact verification around intermediate question–answer (QA) pairs has recently emerged as a means to improve transparency and interpretability. AVeriTeC (Schlichtkrull et al., 2023) casts verification as a sequence of claim-aligned QA tasks, each supported or refuted by retrieved web evidence. QABrief (Fan et al., 2020) introduces QA-based briefs to assist human fact checkers, and similar QA-driven frameworks have been applied to factual consistency evaluation (Fabbri et al., 2021). However, most current pipelines (e.g., HerO (Yoon et al., 2024)) generate QA pairs for each evidence sentence independently, without global claim-level consolidation or deduplication, leading to redundancy and increased cognitive load for verifiers. Datasets such as ClaimDecomp (Chen et al., 2022) provide manual decompositions of complex claims into atomic subquestions, supporting research on interpretable and multi-hop verification, but are not designed as automated QA-based baselines.

Evaluation Metrics. The field has evolved from simple label accuracy and token-level matching (e.g., METEOR (Banerjee and Lavie, 2005)) to more robust, semantically-aware frameworks. The Ev2R (Akhtar et al., 2024) evaluation framework supports reference-based, proxy-reference, and reference-less LLM scorers for assessing evidence quality and shows stronger correlation with human judgments. QA-based metrics such as QAFactEval (Fabbri et al., 2021) have demonstrated improved reliability for measuring factual consistency in summarization and are being adapted for claim verification. Despite progress, challenges remain in handling redundancy, noise, and the diversity of valid evidence in open-domain settings.

3 Method

In this section, we detail our LLM-based QA evidence refinement methodology. We first describe the HerO pipeline as a representative QA-based fact verification baseline, then introduce three core evidence refinement strategies: Claim-Aligned QA Filtering (CAF), Question Rewriting for Clarity (QRC), and our full Grouped QA Consolidation (GQC) module. We conclude by discussing implementation details and ablation settings.

3.1 System Overview

Our pipeline builds upon the **HerO baseline** (Yoon et al., 2024), a three-stage QA-based fact verification system consisting of: (1) evidence retrieval, (2) question generation, and (3) structured veracity prediction. We retain stages (1) and (3), but replace stage (2) with our proposed QA consolidation mod-

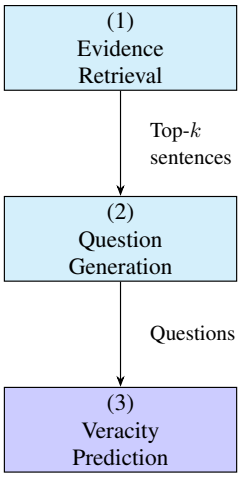

Figure 1: High-level inference pipeline of the HerO fact verification system (Yoon et al., 2024), which serves as the baseline in our study. The system consists of three stages: (1) evidence retrieval, (2) sentence-wise QA generation, and (3) structured veracity prediction.

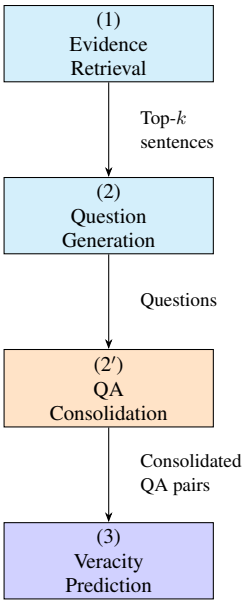

Figure 2: End-to-end inference pipeline with QA consolidation (Step 2′). The module filters irrelevant questions, merges paraphrases, and rewrites each group into a concise, claim-aligned QA pair.

ule that filters, clusters, and rewrites QA pairs at the claim level. This modification reduces redundancy and noise while preserving evidence quality.

As illustrated in Figure 1, the HerO pipeline operates as follows. Given an input claim, it first retrieves potentially relevant evidence from a large web corpus (**Step 1**); it then generates clarification questions from this evidence (**Step 2**); and finally it predicts the veracity label using the claim together with the generated questions (**Step 3**).

Concretely, (1) during evidence retrieval a *frozen* LLaMA-3.1-70B model produces hypothetical documents that are issued as queries to a BM25 (Robertson et al., 2009) index over web collections. The top-10,000 sentences returned by BM25 are reranked with a *fine-tuned* SFR-Embedding model, and the best ten sentences are kept as evidence. (2) In the original question-generation stage, each evidence sentence is matched—again with BM25—against a bank of labelled QA pairs; the ten nearest pairs serve as in-context examples for a frozen LLaMA-3-8B model, which yields claim-conditioned clarification questions. (3) In the veracity-prediction stage, a fine-tuned LLaMA-3.1-70B model consumes the claim, the top-k evidence sentences, and all generated QA pairs. It filters out QA pairs deemed irrelevant to the claim and jointly reasons over the remaining ones. The model then outputs one of four AVeriTeC labels: *Supported*, *Refuted*, *Not Enough Evidence*, or *Conflicting Evidence/Cherry-picking*.

3.2 LLM-based QA Evidence Refinement

The baseline pipeline (see 3.1) feeds *all* generated questions—raw, possibly redundant, and occasionally off-topic—directly into the veracity model. We observe two systematic weaknesses:

1. **Intra-claim redundancy.**
   A single claim often triggers several near-duplicate questions (e.g., "When was he born?" vs. "What is his date of birth?"). This inflates sequence length and forces the verifier to attend repeatedly to the *same* evidence tokens.

2. **Semantic drift.**
   Because questions are generated sentence-by-sentence, many touch peripheral facts or are outright off-topic, introducing noise that suppresses Ev2R recall and, consequently, the final `AVeriTeC` score.

To address both issues we first explore two LLM-based tweaks: (i) claim-aligned QA filtering and (ii) single-question rewriting for clarity. Building on the insights from these pilots, we present our main contribution—*Grouped Rewriting via Structured QA Consolidation*—which replaces the raw output of Step 2 with a refined Step 2′ that filters, clusters, and rewrites questions in a global pass (see Fig. 2).

(1) Claim-Aligned QA Filtering (CAF). We first discard QA pairs that are semantically unrelated to the claim. Given a claim c and QA pair (q_i, a_i), a frozen `Llama-3.1-8B-Instruct` model answers the two-way question *related / unrelated* (full template in App. A.1). Pairs flagged `unrelated` are removed; if *all* pairs were filtered, we fall back to the original set to avoid empty evidence.

(2) Question Rewriting for Clarity (QRC). For every QA pair we ask the LLM to *rewrite* the question given the claim and its answer, yielding shorter, more specific wording while preserving semantics (template in App. A.2).

(3) Grouped Rewriting via *Structured QA Consolidation* (GQC). Steps (1) and (2) treat QA pairs *independently*. Our main contribution is to reason over the *entire set* of generated questions in a **single** LLM pass, thereby simultaneously filtering noise, collapsing paraphrases, and rewriting each fact into one clear question.

Motivation. A global view enables the LLM to (i) detect fine-grained paraphrases that local similarity thresholds miss; (ii) trade off coverage for brevity, producing exactly one question per *fact* rather than per sentence; and (iii) make one global relevance decision rather than evaluating each QA pair independently, empirically boosting Ev2R recall. Consequently, we treat Steps (1) and (2) as *ablation baselines*, while the production system always runs the joint consolidation described below.

Step 1 — Grouping & Filtering. The LLM receives *all* questions for a claim and must

- mark indices of genuinely off-topic questions (`irrelevant`);
- partition the remainder into `groups` whose members can be answered by the *same* evidence sentence.

Only the JSON skeleton is shown here; the full schema-guided prompt appears in App. A.2.

> "Given *claim* + numbered *questions*, return `{"groups": [...], "irrelevant": [...]}`. Every question index must appear *exactly once*."

Step 2 — Rewriting each Group. Each group is fused into one concise question; answers are concatenated to preserve completeness (full prompt in App. A.4).

Algorithm 1 Grouped QA Consolidation

Require: Claim c, QA set $\{(q_i, a_i)\}_{i=1}^{N}$

(groups, irrelevant) $\leftarrow$ LLM_GROUP$(c, \{q_i\})$
new $\leftarrow [\,]$
for all $g \in$ groups **do**
  $Q_g \leftarrow \{q_i \mid i \in g\}$; $A_g \leftarrow \{a_i \mid i \in g\}$
  $\hat{q} \leftarrow$ LLM_REPHRASEGROUP(c, Q_g)
  $\hat{a} \leftarrow$ JOIN(A_g)
  new $\leftarrow$ new $\cup \{(\hat{q}, \hat{a})\}$
end for
return new

Pseudocode. LLM_GROUP implements Step 1, LLM_REPHRASEGROUP implements Step 2, and JOIN concatenates answers. Alg. 1 summarises the workflow.

Discussion. Joint consolidation delivers three qualitative advantages:

1. **Redundancy reduction.**
   Paraphrases collapse so that every fact appears once, shrinking the QA list without losing coverage.
2. **Noise suppression.**
   Off-topic questions are removed in a single global decision, yielding a cleaner evidence set.
3. **Improved clarity.**
   Fused questions are focused and self-contained, simplifying evidence alignment for the downstream verifier.

3.3 Limitations of Existing Metrics

Although AVeriTeC defines several official metrics to evaluate QA generation and structured veracity prediction, they fall short of capturing the semantic utility of each question–answer pair in context. Below, we outline the core limitations:

- **Ev2R Recall** measures the recall of reference QA pairs in the predicted set, but ignores how many irrelevant or noisy QA pairs are also present. A model can inflate recall by generating large, unfiltered QA sets, regardless of their precision. Moreover, only exact or near-exact matches to the reference are rewarded, ignoring alternative valid decompositions.
- **New AVeriTeC Score** imposes a hard cutoff over Ev2R recall. If a submission falls below a fixed threshold ($\lambda = 0.25$), its veracity

prediction is ignored (scored zero), regardless of partial validity. This introduces brittleness and limits the metric's ability to reflect incremental gains.

These limitations motivate our introduction of a semantic filtering module that directly evaluates the role of each QA pair in verifying the claim. Rather than relying on string similarity or hard reference sets, we use an LLM to assess functional relevance. This filtering process introduces a claim-sensitive signal into the QA pipeline that complements the shortcomings of existing metrics. Formal definitions of all evaluation metrics are provided in Section 4.2.

4 Experiments

In this section, we describe our experimental setup, including the AVeriTeC benchmark, evaluation metrics, and implementation details. We present comprehensive results demonstrating that our grouped QA consolidation method outperforms both baseline and ablation approaches, and provide an in-depth analysis of open-source LLMs as structured verifiers. Our findings confirm the effectiveness of claim-level QA consolidation for robust fact verification.

4.1 Benchmark Setup

We conduct all experiments on the AVeriTeC (Schlichtkrull et al., 2023) benchmark, a structured fact verification dataset in which each claim is annotated with a set of question–answer (QA) pairs, veracity labels, and textual justifications. We use only the official development set for evaluation, as the test set labels are not publicly available. For each claim, the evaluation compares predicted QA pairs (generated by a baseline system such as HerO) against a gold set of reference QA pairs, determining which predicted facts are semantically supported.

Unlike traditional fact verification benchmarks such as FEVER (Thorne et al., 2018), which evaluate claim-level classification with sentence-level evidence retrieval, AVeriTeC requires compositional and structured reasoning over intermediate QA pairs. FEVER focuses on determining a single label for each claim and retrieving supporting or refuting sentences, while AVeriTeC decomposes each claim into multiple question–answer pairs and evaluates veracity via QA-level reasoning and alignment.

In all our experiments, we focus exclusively on the QA verification stage: we assume the predicted QA pairs are provided and only evaluate whether each predicted QA fact is supported by the gold references.

4.2 Evaluation Metrics

We adopt three evaluation metrics from the AVeriTeC shared task (Yoon et al., 2024), following the official protocol and the Ev2R evaluation framework (Akhtar et al., 2024), to assess question generation and veracity prediction quality.

Q-only Ev2R. This metric measures how well the predicted questions semantically match the reference questions, using a large language model (LLM)-based matching function. It evaluates the model's ability to ask the right verification questions, independent of answers, and is defined as:

$$\text{Q-only Ev2R} = \frac{\text{\# Matched Reference Questions}}{\text{\# Total Reference Questions}}$$

Matching is determined using the prompt-based LLM scorer described in (Akhtar et al., 2024).

Q+A Ev2R. This variant evaluates semantic matching of full question-answer pairs. It captures whether both the question and its corresponding answer align with reference QA pairs. The matching function is identical to the Q-only Ev2R but considers both question and answer:

$$\text{Q+A Ev2R} = \frac{\text{\# Matched Reference QA Pairs}}{\text{\# Total Reference QA Pairs}}$$

AVeriTeC Score. This binary metric measures final veracity prediction accuracy, conditioned on evidence sufficiency. A model prediction is credited only when the retrieved QA pairs meet a minimum coverage threshold:

$$\text{AVeriTeCScore} = \begin{cases} \text{VeracityAccuracy}, & \text{if Q+A Ev2R} \geq \lambda \\ 0, & \text{otherwise} \end{cases}$$

where $\lambda = 0.25$ is a fixed threshold. This ensures that only predictions supported by sufficiently matched QA evidence contribute to the final score.

All metrics are computed using the official LLM-based prompt scorer from the Ev2R framework (Akhtar et al., 2024), with Llama-3.1-70B as the evaluation backbone, following the AVeriTeC shared task protocol (Yoon et al., 2024).

Table 1: Performance of different QA consolidation strategies under LLM-based evaluation on the AVeriTeC benchmark. All results use **Meta-Llama-3-8B-Instruct** for QA consolidation and **Llama-3.1-70B** as the downstream verifier. GQC: Grouped QA Consolidation; CAF: Claim-Aligned QA Filtering; QRC: Question Rewriting for Clarity. The best score in each column is shown in bold.

Method	Q-only Ev2R	Q+A Ev2R	AVeriTeC Score
HerO	**0.757**	0.540	0.278
GQC	0.753	**0.566**	**0.312**
CAF	0.730	0.553	0.278
QRC	0.498	0.434	0.216

4.3 Experimental Results

Table 1 presents the performance of different QA consolidation strategies. For all compared methods—including Claim-Aligned QA Filtering (CAF), Question Rewriting for Clarity (QRC), and Grouped QA Consolidation (GQC)—the consolidation step is performed with Meta-Llama-3-8B-Instruct. Final veracity prediction is evaluated using a fixed Llama-3.1-70B verifier. The key evaluation metric is the AVeriTeC Score, which measures final fact verification accuracy *conditioned on evidence sufficiency*: a model's prediction is only credited if the retrieved QA pairs achieve a minimum Q+A Ev2R coverage threshold ($\lambda = 0.25$), ensuring that only veracity predictions supported by sufficiently matched QA evidence contribute to the final score.

The baseline HerO system achieves strong Q-only Ev2R (0.7574) and Q+A Ev2R (0.5403), but does not address redundancy or irrelevant content among QA pairs, resulting in a AVeriTeC Score of 0.278. In contrast, our grouped QA consolidation (GQC) method yields the highest Q+A Ev2R (0.5664) and achieves a substantial improvement in AVeriTeC Score (from 0.278 to 0.312, a 12.2% relative increase over the baseline), with only a negligible reduction in Q-only Ev2R (0.7526). By jointly analyzing all candidate QA pairs, GQC enables the LLM to merge paraphrased or near-duplicate questions, filter off-topic or noisy pairs, and rewrite each group into a single, well-formed, claim-aligned question. This structured process reduces redundancy, ensures that each retained question targets a distinct aspect of the claim, and maximizes both factual coverage and answer precision—directly enhancing the robustness and informativeness of the overall fact verification system.

To illustrate the effect of grouped QA consolidation, consider the following real example from our evaluation set. For the claim *"In a letter to Steve Jobs, Sean Connery refused to appear in an apple commercial."*, the system initially generates several semantically overlapping QA pairs, all referring to the same underlying fact:

Example Claim:

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

Representative original QA pairs:

- **Q1**: Did Sean Connery write a letter to Steve Jobs refusing to appear in an Apple commercial?
- **Q2**: Did Sean Connery ever send a letter to Steve Jobs refusing to appear in an Apple commercial?
- **Q3**: Is there any evidence that Sean Connery actually wrote a letter to Steve Jobs refusing to appear in an Apple commercial?

Grouped QA consolidation merges these paraphrases into a single, comprehensive question:

After consolidation:

- Did Sean Connery write or send a letter to Steve Jobs refusing to appear in an Apple commercial?

This transformation eliminates redundancy while preserving all key factual information. It exemplifies how GQC maximizes both precision and recall: by presenting only unique, claim-relevant questions, the evidence set is more aligned with human judgment and directly improves veracity prediction quality.

Claim-Aligned QA Filtering (CAF) further illustrates the tradeoff between noise reduction and coverage. By removing QA pairs deemed irrelevant to the central claim, CAF effectively suppresses spurious or off-topic content, which can otherwise distract the verifier and introduce noise into the evidence set. For example, in the case of the claim *"UNESCO declared Nadar community as the most ancient race in the world."*, CAF filters out the following question as unrelated:

Example Claim:

"UNESCO declared Nadar community as the most ancient race in the world."

QA pair filtered out by CAF:

- What is the current social status of the Nadar community in Tamil Nadu?

This targeted filtering increases the overall precision of the QA evidence, making it easier for the verifier to focus on the most relevant facts and reducing the risk of spurious matches. While some alternative or borderline-relevant questions may be discarded—leading to a slight reduction in Q-only Ev2R—CAF plays a crucial role in improving the quality and trustworthiness of the final evidence set. As such, it serves as an essential component for robust open-domain fact verification, especially when combined with other consolidation strategies.

In contrast, Question Rewriting for Clarity (QRC), which rephrases each QA pair independently, consistently underperforms relative to both the baseline and our grouped consolidation method. Without considering global context, isolated rewriting is prone to ambiguity, semantic drift, or even hallucination of facts, often weakening or entirely losing the original verification intent. This issue is exemplified by the following case:

Example Claim:

"UNESCO declared Nadar community as the most ancient race in the world."

Original Question:

- Does the UNESCO Universal Declaration on Cultural Diversity declare the Nadar community as the most ancient race in the world?

After QRC rewriting:

- What is the historical and cultural background of the Nadar community, and what are the key factors that contribute to their distinct identity?

(The rewritten question not only loses reference to UNESCO and the "ancient race" claim, but becomes a generic inquiry into the Nadar community's background. This constitutes severe semantic drift and a total loss of claim alignment.)

Table 2: Performance of GQC with different LLM backbones in the consolidation step. All results are evaluated with a fixed Llama-3.1-70B verifier.

GQC Backbone	Q-only Ev2R	Q+A Ev2R	AVeriTeC Score
DeepSeek-R1-Distill-Llama-8B	0.739	0.548	0.294
Llama-3.1-8B-Instruct	0.753	0.566	0.312
Qwen2.5-7B-Instruct	0.766	0.579	0.318
Qwen2.5-32B-Instruct	**0.771**	**0.582**	**0.327**

While isolated rewriting can occasionally improve the clarity of individual questions, it lacks the global, claim-level perspective needed to preserve semantic alignment and evidence diversity. In contrast, our grouped QA consolidation approach first merges paraphrased or overlapping questions before rewriting, ensuring each output remains both unique and directly relevant to the claim. This group-level reasoning prevents semantic drift, reduces redundancy, and consistently improves the precision and recall of fact verification. Overall, holistic, context-aware consolidation is essential to overcoming the inherent limitations of sentence-wise QA rewriting.

Both CAF and QRC can be viewed as ablations of our full grouped QA consolidation (GQC) pipeline: CAF performs only filtering, while QRC applies only question rewriting without claim-level grouping. Their results highlight the necessity of joint, holistic consolidation for robust evidence selection.

Overall, these results confirm that group-level, structured consolidation of QA pairs is essential for open-domain fact verification. Our approach not only increases Q+A Ev2R by 4.8% absolute (from 0.5403 to 0.5664) but also delivers a notable 12.2% improvement in the AVeriTeC Score, while maintaining high Q-only Ev2R. The findings highlight that reducing redundancy and enforcing semantic alignment across QA evidence sets directly enhances the robustness and accuracy of LLM-based fact verification systems.

4.4 Ablation Study: GQC Backbone Analysis

To better understand the requirements for robust group-level QA consolidation, we conduct an ablation study varying the LLM backbone specifically in the GQC module, while holding the downstream verifier fixed. Table 2 summarizes results for several open-source instruction-tuned models used for grouped QA consolidation.

The results show that all large instruction-tuned models benefit from grouped QA consolidation, but the best overall performance is achieved with

the Qwen2.5-32B-Instruct backbone. Notably, Qwen/Qwen2.5-7B-Instruct outperforms Llama-3.1-8B-Instruct across all metrics, while Qwen2.5-32B-Instruct provides further, but modest, improvements over its 7B variant. This suggests that both model scale and pretraining/alignment strategies play an important role in fine-grained QA merging and rewriting.

Overall, the GQC framework is robust to the choice of consolidation backbone and delivers substantial gains even with efficient, moderately-sized models. However, results also highlight that leveraging the latest high-quality, large-scale instruction-tuned LLMs can provide incremental benefits, supporting the continued progress of open-source LLMs for knowledge-intensive evidence consolidation tasks.

4.5 Open LLMs as Ev2R Scorers

We benchmark five instruction-tuned open models—QWEN2.5-7B/14B/32B and LLAMA-3.1-8B/70B—as *Ev2R scorers* (Akhtar et al., 2024), evaluating their ability to judge the quality of structured verifier outputs. Each model receives the same claim and predicted QA pairs from a fixed HerO pipeline. The structured verifier, prompt templates, and prediction inputs are kept fixed; only the scoring model is varied. Table 3 summarizes results across Q-only, Q+A, and AVeriTeC metrics.

Model	Q-only Ev2R	Q+A Ev2R	AVeriTeC Score
`Qwen2.5-7B`	0.000	0.100	0.000
`LLaMA3.1-8B`	0.000	0.100	0.000
`Qwen2.5-14B`	0.358	0.501	0.246
`Qwen2.5-32B`	0.715	0.521	0.254
`LLaMA3.1-70B`	**0.753**	**0.566**	**0.312**

Table 3: Performance of different LLMs used as Ev2R scorers. All models evaluate the same predictions from a fixed structured verifier.

Analysis. Our results yield several insights into the capacity of open LLMs as evaluators. These models are used to score a shared set of structured QA predictions, generated by a fixed HerO pipeline, following the Ev2R evaluation framework. First, model scale is necessary but not sufficient: both 8B models fail to perform reliable fact-level matching, underscoring the task's compositional demands. Qwen2.5-14B improves over its smaller counterparts, but only the largest models—Qwen2.5-32B and LLaMA-3.1-70B—achieve robust performance across all metrics. Notably, LLaMA-3.1-70B sets a new ceiling for open models, reaching 0.753 Q-only Ev2R and 0.312 AVeriTeC Score without task-specific tuning.

Yet challenges remain. Even strong models are brittle to minor format violations (e.g., JSON malformation) and highly sensitive to upstream QA quality. Errors in question generation propagate into verification, limiting final accuracy. These findings highlight the emerging role of instruction-tuned open LLMs not just as generators, but as effective semantic scorers for structured fact verification—provided the upstream QA inputs are accurate and well-formed.

5 Conclusion

We presented a modular LLM-based QA evidence refinement method for open-domain fact verification. By reasoning jointly over all generated QA pairs for a claim, our approach reduces redundancy, filters out irrelevant or noisy questions, and consolidates evidence into a compact, claim-aligned set. Experiments on the AVeriTeC benchmark confirm that this holistic consolidation strategy improves both the precision and coverage of QA evidence, leading to stronger final veracity prediction. Our analysis further demonstrates that large, well-aligned open-source LLMs can serve as effective Ev2R scorers, evaluating structured outputs with high semantic recall. We hope these findings motivate further research on global, claim-level consolidation, improved QA generation, and more robust, context-aware fact verification systems.

Acknowledgements

The work was partially supported by EU Projects Graph Massivizer (GA 101093202), enRichMyData (GA 101070284), and SMARTY (GA 101140087), and the Research Council of Norway through its Centres of Excellence scheme, Integreat—Norwegian Centre for knowledge-driven machine learning, project 332645.

References

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

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. *arXiv preprint arXiv:2011.05448*.

Yuan He, Bailan He, Zifeng Ding, Alisia Lupidi, Yuqicheng Zhu, Shuo Chen, Caiqi Zhang, Jiaoyan Chen, Yunpu Ma, Volker Tresp, and 1 others. 2025. Supposedly equivalent facts that aren't? entity frequency in pre-training induces asymmetry in llms. *arXiv preprint arXiv:2503.22362*.

Nico Potyka, Yuqicheng Zhu, Yunjie He, Evgeny Kharlamov, and Steffen Staab. 2024. Robust knowledge extraction from large language models using social choice theory. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 1593–1601.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. Hero at averitec: The herd of open large language models for verifying real-world claims. *arXiv preprint arXiv:2410.12377*.

Dongzhuoran Zhou, Yuqicheng Zhu, Yuan He, Jiaoyan Chen, Evgeny Kharlamov, and Steffen Staab. 2025. Evaluating knowledge graph based retrieval augmented generation methods under knowledge incompleteness. *arXiv preprint arXiv:2504.05163*.

Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, and Steffen Staab. 2025. Argrag: Explainable retrieval augmented generation using quantitative bipolar argumentation. In *International Conference on Neural-Symbolic Learning and Reasoning*. To appear.

A Prompt Templates and Implementation Details

A.1 Claim–Aligned QA Filtering

System instruction. You are given a claim and a question–answer pair. Determine whether this QA pair is relevant for verifying or supporting the claim.
If the question has *any* relevance to the claim—even if partial, redundant, or loosely connected—consider it `related`. Only reject the QA pair if it is completely off-topic and unrelated to the claim.
Respond with a *single* word: either "**related**" or "**unrelated**".

Claim: {claim}
Question: {question}
Answer: {answer}

Response:

This prompt is fed to a frozen `Llama-3.1-8B-Instruct` model with `temperature` $= 0.0$ to obtain a hard "related / unrelated" decision (Sec. 3.2).

A.2 Question Rewriting for Clarity

Instruction. Improve the following question based on the *claim* and its *answer*. Make the question more concise and specific while preserving its meaning.

Claim: {claim}
Q: {question}
A: {answer}

Improved Question:

All retained questions are rewritten in parallel with `temperature`=0.6 using the same frozen `Llama-3.1-8B-Instruct` backbone.

A.3 Joint Grouping & Filtering Prompt

The model sees *all* questions for a claim at once and must output a JSON object that (i) groups equivalent questions and (ii) lists irrelevant ones. Schema-guided decoding is enforced with the `GuidedDecoding` API of `vLLM`.

You are given a **claim** and a list of **numbered questions**. Your tasks:

1. Identify which questions are unrelated to the claim. Return their indices in a list called "`irrelevant`".
2. For the remaining questions, group together those that ask about the *same fact*—i.e. they can be answered by the *same sentence of evidence*. Return these as an array of objects:

```
"groups": [{"questions": [1, 2, 4]},
           {"questions": [3, 5]}]
```

Constraints

- Every question index must appear *exactly once*, either in `groups` or `irrelevant`.
- Return only the JSON object; do *not* include explanations.

```
Claim: {claim}
Questions:
1. ...
2. ...
```

Now return JSON:

A.4 Group-Level Rephrasing Prompt

Instruction. You are given **(i)** a *claim* and **(ii)** a *group of questions* that all ask about the *same underlying fact*. **Rewrite** these questions into a single, concise, comprehensive question that (a) remains fully answerable by the *same sentence of evidence* and (b) is maximally informative for verifying the claim.

Claim: {claim}

Questions (paraphrases of the same fact):

```
1. ...
2. ...
```

Output format (only one line):

```
Rephrased question:
<your single fused question here>
```

Guidelines:

- Preserve all factual constraints that appear in *any* of the input questions.
- Remove redundant words, vague pronouns, or rhetorical flourishes.
- Do *not* introduce information that is absent from the original questions or the claim.
- Keep the wording as short as possible while staying precise.

Generation settings. We pass the above template to the frozen `Llama-3.1-8B-Instruct` model with `temperature = 0.3`. Only the fused question is kept; answers inside the same group are

concatenated verbatim, as described in Sec. 3.2.