Mask-to-Correct^{+}: Leveraging Retriever Diversity for Masking-guided Faithful Fact Correction

Payel Santra[*], **Lavisha Sharma**[*], **Madhusudan Ghosh** and **Partha Basuchowdhuri**
Indian Association for the Cultivation of Science, Jadavpur, Kolkata-700032, India
{payel.iacs, lavishasharma2612, madhusuda.iacs}@gmail.com
and partha.basuchowdhuri@iacs.res.in

Abstract

The rapid spread of misinformation on social media highlights the need for robust, automated fact correction frameworks. However, existing works rely on supervised learning from manually annotated claim-evidence pairs, which are scarce and prone to biases, limiting their generalization across domains. Moreover, these methods overlook semantic faithfulness in their correction process. To address these challenges, we propose **Mask-to-Correct ($\mathbf{M_2C}$)**, a training-free, inference-only Retrieval Augmented Generation (RAG) based framework that leverages *diversity-aware masking* to identify erroneous spans of claims and evaluate the faithfulness of corrections using retrieved evidence. However, the effectiveness of RAG heavily depends on the choice of retriever, which may vary across queries. To mitigate this, we further introduce $\mathbf{M_2C^{+}}$, an ensemble-based framework that combines corrections across multiple rankers to reduce retrieval bias and improve robustness. Extensive experiments on the benchmark datasets demonstrate that our proposed frameworks consistently outperform all baselines, achieving up to 14% improvement in SARI scores, without using gold evidence.

1 Introduction

Large language models (LLMs) have made a significant impact across a wide range of tasks (Sun et al., 2023; Fu et al., 2023; Yue et al., 2024). However, LLMs (Touvron et al., 2023; Wang and Komatsuzaki, 2022) have intensified the generation of misinformation with the help of suitably crafted adversarial prompts (Zou et al., 2023). The topically coherent and fluent nature of LLM-generated text (Liu et al., 2021b) potentially makes it even harder to detect any injected misinformation (Parry et al., 2024). On the other hand, the inherent susceptibility of language models towards hallucinations (Ji et al., 2023) often results in factually inconsistent outputs.

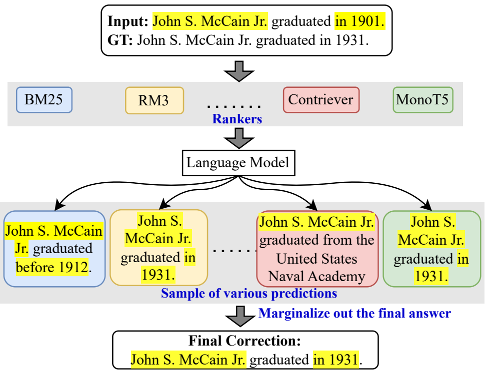


Figure 1: An illustrating example of our proposed model. Given an incorrect claim, each retriever yields a different set of evidence, leading to diverse corrections. Our approach encompasses these corrections to marginalize out a factually correct version of the claim. Here, 'GT' denotes the ground-truth claim.

In this paper, we primarily address the task of factual error correction, which involves editing a given input sentence (or claim) to make it factually consistent with the supporting evidence. Unlike the fact verification task, which involves classifying a claim's veracity class ("TRUE", "FALSE", etc.) (Schuster et al., 2019, 2021; Jiang et al., 2021) based on supporting or refuting evidence (Asai et al., 2023), fact correction extends this task by revising incorrect claims into correct ones while preserving their semantics. Most standard approaches of this task focus on preserving grammar and fluency using fully or distantly supervised approaches (Shah et al., 2020; Thorne and Vlachos, 2021a; He et al., 2024), but ignore faithfulness[1] to the supporting evidence. Moreover, these methods require annotated evidence of each claim, which

---

[*] These authors contributed equally to this work.

[1] Semantic faithfulness refers to preserving a claim's core meaning, its proposition, structure, and intent while modifying only factually incorrect spans based on external evidence. Unlike grammaticality or fluency, which focus on surface form, semantic faithfulness ensures the integrity of meaning with respect to both the original claim and the supporting evidence.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3808–3825
July 2-7, 2026 ©2026 Association for Computational Linguistics

is resource-intensive and time-consuming and also susceptible to pooling and exposure biases (Yang et al., 2019) (see Appendix C).

Inspired by Huang et al. (2023), we address these challenges by introducing training-free approaches, such as zero-shot prompting followed by a Retrieval-Augmented Generation (RAG)-based framework (Lewis et al., 2020; Izacard and Grave, 2021), which enhances an LLM's response generation by integrating external knowledge from large-scale corpora, thereby eliminating the need for manually annotated ground-truth evidence. However, such prompting approaches face some challenges: **Ensuring minimal change:** The corrected claim should preserve the original semantics. Identifying potentially false or weakly supported spans within a claim may help us address it convincingly. **Retriever quality dependency:** As illustrated in Figure 1, the effectiveness of RAG-based correction depends heavily on the quality of the retrieved evidence. A single retriever may fail to capture diverse or ambiguous contexts and accumulating complementary evidence from multiple retrievers leads to more reliable corrections.

To overcome these challenges, we propose a novel masking aware, three-stage framework M_2C (Mask-to-Correct) that identifies and masks erroneous spans in a claim, generates corrected candidates using retrieved evidence, and selects the most faithful revision through a semantic–factual scoring function. We further introduce M_2C^+, an ensemble-based extension that aggregates multiple retrieval generation pathways for improved robustness. Importantly, to our knowledge, our work is the first to integrate RAG with a diversity-aware masking strategy that ensures broader coverage of plausible error regions while minimizing unnecessary edits, hallucinations, thereby increasing faithfulness. Thus, the novelty of this approach lies in a task-specific, lightweight aggregation across different retrievers that enhances correction stability without additional training overhead. Furthermore, to strengthen our study, we conducted extensive analyses in multiple dimensions, including the impact of different masking strategies, retriever combinations, computational efficiency, and correction scoring schemes, along with detailed sensitivity and ablation studies.

Our contributions are summarized as follows: 1) We introduce a training-free fact correction framework, M_2C, that does not require gold evidence. 2) We propose a diversity-aware masking strategy to identify high-impact erroneous spans more effectively than traditional methods. 3) To the best of our knowledge, we are the first to introduce majority voting-based ensembling approach for the downstream task through our framework M_2C^+.

2 Related Work

Fact Verification. Early works on fact verification explored supervised methods using pre-trained models (Stammbach and Neumann, 2019; Krishna et al., 2022; Soleimani et al., 2020; Chernyavskiy and Ilvovsky, 2019), multitask learning (Hidey and Diab, 2018; Lewis et al., 2020), and Graph-based learning (Zhou et al., 2019; Liu et al., 2020). Recent methods have deployed LLMs with in-context learning (ICL) and RAG to improve scalability and factual grounding (Kojima et al., 2022a; Santra et al., 2024; Min et al., 2022; Santra et al., 2025).

Retrieval Augmented Generation. RAG-based methods integrate retrieval from external KBs to reduce hallucinations during generation (Guu et al., 2020; Shi et al., 2023; Lan et al., 2023; Jiang et al., 2023; Zhang et al., 2023; Izacard and Grave, 2020; Borgeaud et al., 2022). Recent advances focus on improving retrieval through adaptive context selection (Jeong et al., 2024), iterative refinement, and relevance-based integration (Glass et al., 2022). A similar work, EoR (Li et al., 2024), adaptively fuses retrievals from different sources to reduce evidence-level inconsistencies in retrieval-augmented QA task. In contrast, our framework performs correction-level ensembling across independently corrected outputs using various retrievers.

Fact Correction. Existing supervised strategies include methods that target factual inconsistencies in text summarization (Fabbri et al., 2022; Adams et al., 2022), employ masking techniques to guide correction (Shah et al., 2020; Thorne and Vlachos, 2021b), or perform iterative edits based on predicted truthfulness scores (Chen et al., 2023). In some studies, distantly supervised correctors were trained on synthetic data, avoiding the need for explicit masking (He et al., 2024; Ashok et al., 2023). Recent studies have addressed this problem with unsupervised approaches by leveraging a QA-based pipeline (Huang et al., 2023) with knowledge graph-based retrieval (Bayat et al., 2023).

In contrast to previous studies, we adopt a masking-guided unsupervised approach for the fact

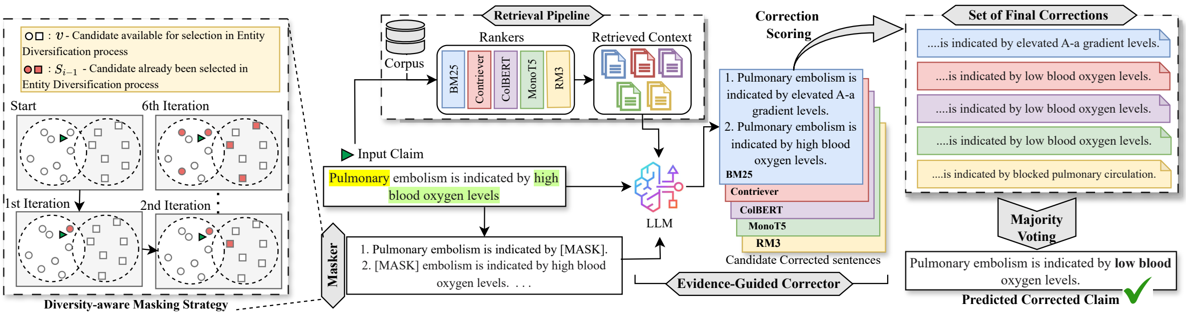


Figure 2: Schematic overview of our proposed M_2C^+. Given a claim, the diversity-aware masker (an iterative module) identifies and prioritizes spans that are likely to be incorrect. It selects entities based on similarity and diversity (e.g., selecting the most similar entity in the 1^{st} iteration (from the same cluster); in the 2^{nd} iteration, it adds an entity that remains relevant to the claim but is diverse from the one previously selected (e.g., from a different semantic cluster); and so on.). Here, we have highlighted top-2 entities for masking. Top documents are then retrieved from an external corpus using multiple retrievers. Each retrieval path is used by an LLM to generate corrected candidates. A correction scoring module selects the best correction per retriever, and majority voting is applied across these outputs to produce the final corrected claim.

correction task by integrating an ensemble-based RAG framework to ensure both faithfulness and factuality. Unlike traditional maskers, we proposed a diversity-aware masking strategy.

3 Methodology

In this paper, we address the task of faithful fact correction, which involves rectifying an incorrect claim using contextual evidence. Formally, given an input claim $\mathbf{x}$ (correct, incorrect or ambiguous), our goal is to design an automated system ϕ that identifies and rectifies factual errors (if they exist) to generate a revised claim $\hat{\mathbf{x}}$ that is consistent with the associated evidence $\mathcal{E}(\mathbf{x})$. More formally, $\phi : (\mathbf{x}, \mathcal{E}(\mathbf{x})) \mapsto \hat{\mathbf{x}}$. We hypothesize that the predicted claim $\hat{\mathbf{x}}$ should be grammatically correct and minimally altered to preserve the original semantics of $\mathbf{x}$. Furthermore, our method is designed to maintain factual consistency if the $\mathbf{x}$ is already correct. We propose a novel three-stage masking-guided, training-free, and annotation-free framework, M_2C, that leverages the reasoning capabilities of LLMs. The key components are described as follows:

3.1 Masker

The objective of this module is to identify salient and diverse spans within a claim that are most likely to require factual revision, while preserving the overall structure and semantics of the input without explicitly verifying correctness. Formally, given an input claim $\mathbf{x} = \{x_i\}_{i=1}^{n}$, where x_i denotes the i^{th} token, the module first extracts a set of candidate entities or spans $E_s(\mathbf{x}) = \{e_i\}_{i=1}^{s}$, with $s < n$. Each entity e_j is then independently masked to produce a set of perturbed claims $\{\tilde{\mathbf{x}}_j\}_{j=1}^{s}$, where $\tilde{\mathbf{x}}_j$ is obtained by replacing e_j with a $[MASK]$ token. However, standard masking strategies such as random masking (Thorne and Vlachos, 2021a) or purely entity based masking (Huang et al., 2023) either lack in semantic focus or rely solely on syntactic boundaries, often missing contextually important spans and leading to suboptimal corrections. To address this limitation, we propose a Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) based **diversity-aware masking paradigm** to iteratively re-rank candidate entities. This yields a prioritized subset $\mathcal{R}_m = \{r_i\}_{i=1}^{m}$, where $m \leq s$, ensuring that the selected spans collectively capture diverse and informative aspects of the claim for our downstream task.

As illustrated in Figure 2, using MMR algorithm, we iteratively add those entities in an incrementally growing list that are both relevant to the claim and diverse with respect to one another. More formally, it is a function of the following: (a) the iteration step $i < m$, (m - the number of entities to be selected), (b) the set of entities selected until the $(i-1)^{th}$ step, $S_{i-1} \subseteq \mathcal{R}_m$, and (c) a hyperparameter $\alpha \in [0, 1]$ that balances relevance with diversity. Formally, $\forall v \in \{E_s \setminus S_{i-1}\}$,

$$\theta_{\text{MMR}}(\mathbf{x}, v, S_{i-1}; \alpha) = \alpha\, \theta(\mathbf{x}, v) - (1-\alpha) \max_{s \in S_{i-1}} \theta(v, s), \tag{1}$$

where, θ is the similarity function. For the claim mentioned in Figure 2, the method first selects the most relevant one (e.g., ‘ high blood oxygen levels’), and then selects diverse but relevant entities (like ‘Pulmonary’) by ignoring those similar to the already selected ones.

3.2 Evidence-Guided Corrector

Given the set of masked claims $\{\tilde{\mathbf{x}}_j\}_{j=1}^{k}$ derived from an input claim $\mathbf{x}$, this module performs evidence-guided correction through a RAG based approach. Intuitively, the masking step isolates uncertain spans, but without external grounding, directly reconstructing them risks hallucination or semantic drift. To mitigate this, we shift from purely generative reconstruction to an evidence based generation process, where corrections are explicitly anchored in retrieved knowledge. Here, the model directly leverages a small set of examples which are similar to the current claim from an external corpus $\mathcal{T}$, eventually including these with the masked claim as a part of an input prompt to an LLM (Liu et al., 2021a; Agrawal et al., 2022; Huang et al., 2022). To retrieve the most relevant examples, we employ a retrieval model $\theta \in \Theta$, which selects top-p similar sentences as evidences based on lexical or semantic similarity. Formally,

$$\phi_{\text{LLM}}(\mathbf{x}, \tilde{\mathbf{x}}_j, \mathcal{E}_{p;\theta}(\mathbf{x})) \mapsto \mathbf{x}'_j, \quad (2)$$

where, $\mathcal{E}_{p;\theta}(\mathbf{x}) \subset \mathcal{T}$ and $\mathbf{x}'_j$ is the predicted corrected sentence of the masked sentence $\tilde{\mathbf{x}}_j$ generated by an LLM. For details regarding our prompt template, please refer to Figure 5 of Appendix G.

3.3 Correction Scoring

In this module, given a set of generated candidate corrected sentences $\{\mathbf{x}'_j\}_{j=1}^{k}$, appended with input claim $\mathbf{x}$, our objective is to identify the most factually consistent and semantically faithful candidate correction $\hat{\mathbf{x}}$ for a given input claim $\mathbf{x}$. We utilize a scoring function $\mathcal{F}(\mathbf{x}'_j)$ by using two metrics as used in (Huang et al., 2023): 1) DocNLI (Yin et al., 2021), which measures factual entailment with the provided evidence $\mathcal{E}_{p;\theta}(\mathbf{x})$, and 2) ROUGE-L (Lin, 2004), which captures the longest common subsequences between the input claim $\mathbf{x}$ and the candidate $\mathbf{x}'_j$, thereby ensuring minimal deviation from the input claim. The candidate having the maximum score is picked as the final corrected claim, $\hat{\mathbf{x}}$. Formally speaking, $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}'_j} \mathcal{F}(\mathbf{x}'_j)$, where,

$$\mathcal{F}(\mathbf{x}'_j) = \lambda \cdot \text{DocNLI}(\mathbf{x}'_j, \mathcal{E}) + (1-\lambda) \cdot \text{ROUGE-L}(\mathbf{x}, \mathbf{x}'_j).$$

The parameter $\lambda \in [0, 1]$ balances factuality and faithfulness. We selected the optimal λ using FEVER validation data (see Appendix F). The schematic diagram of M_2C is depicted in Figure 2.

3.4 Ensemble Based Model

We further extend our framework with a novel ensemble mechanism that aggregates corrected outputs derived from multiple retrieval perspectives to improve robustness and reliability. Intuitively, while a single retriever may capture only a partial or biased view of the evidence space, different retrievers $\theta_i \in \Theta$ (for $i = 1, \ldots, m$) often surface complementary contexts, leading to diverse yet plausible corrections for a given masked input. Instead of committing to any single retrieval signal, we shift to a consensus driven paradigm, where M_2C generates a set of candidate corrections, one per retriever and consolidates them via a hard majority voting strategy. Concretely, the final corrected claim is selected as the candidate receiving the highest number of votes across all retrievers, effectively approximating a form of collective agreement over independently grounded generations (see Figure 2). This design mitigates retriever-specific biases, enhances stability under retrieval noise, and eliminates the need for additional supervision or training. We named this Mask-to-Correct$^+$ or, M_2C^+ [2].

4 Experiment Setup

4.1 Research Questions

The effectiveness of RAG depends on both the quality of the retrieval and the masking strategy used. To understand their impact on downstream performance, we investigate the following research questions: a) **RQ-1**: Does M_2C^+ perform better than the individual retriever-based approach, M_2C? b) **RQ-2**: Can diversity-aware masking strategies lead to improved downstream performance? c) **RQ-3**: How does performance with retrieved evidence differ compared to the gold-standard evidence? d) **RQ-4**: How strongly does retrieval effectiveness correlate with downstream gains?

4.2 Dataset Description

We conduct our experiments on two datasets repurposed for our task: 1) FEVER (Thorne and Vlachos, 2021c), comprising of general-domain claims 2) SciFact (Wadden et al., 2020), consisting of scientific claims. For FEVER we retrieve relevant evidence from the Wikipedia'18 dump, and for SciFact we use the S2ORC corpus (Wadden et al., 2020). Table 1 summarizes these dataset statistics. Detailed description can be found in Appendix B.

4.3 Rankers Investigated

In this work, we employ RAG in the evidence-guided corrector module and analyze the impact of

---

[2] The code is publicly available at: `https://github.com/payelsantra/MaskToCorrect.git`

Dataset	Usage	#Sup	#Ref	#Claims
FEVER	Validation	414	601	1555
	Test	1,593	2,289	3,882
SciFact		43	57	100

Table 1: Statistics of FEVER and SciFact dataset for our downstream fact-correction task.

context selection. We experiment with four single-stage retrievers and one multi-stage ranker. We consider a diverse set of widely used, training-free retrievers and rerankers to mitigate retriever-specific biases and ensure a fair comparison. For consistency, we retrieve the top 50 candidates for each claim across all models.

Single-stage Ranker. We explore two *sparse retrievers*: 1) **BM25** (Robertson and Zaragoza, 2009), a traditional lexical retriever that relies on exact term matching, utilizing a TF-IDF variant. In our experiments, we set the BM25 parameters k_1 and b to 0.9 and 0.4, respectively. 2) **RM3** (Han et al., 2021), a pseudo-relevance feedback method (built on top of lexical retrievers like BM25) to expand the queries and re-rank the retrieved documents.

We also experimented with two *dense end-to-end rankers*: 1) **Contriever** (Lei et al., 2023), a bi-encoder framework fine-tuned on the MS MARCO passage dataset to generate dense embeddings and capture semantic similarity. 2) **ColBERT** (Khattab and Zaharia, 2020), adopts TCT-ColBERT, a late interaction model, to efficiently capture fine-grained token-level interactions.

Two-stage Ranker. In our *retrieve-and-rerank* approach, we utilize **MonoT5** (Pradeep et al., 2021), which uses BM25 to retrieve a small candidate set, which is then refined by a T5-based MonoT5 (a cross-encoder model) re-ranker.

In this paper, our retriever selection covers lexical, pseudo-relevance feedback, dense (bi-encoder, late-interaction), and cross-encoder reranker paradigms. By relying mostly on CPU-compatible methods, our framework remains resource-efficient while enabling robust analysis of retriever variability.

4.4 Maskers Investigated

We compare our proposed masking strategy with the following baseline maskers:
(a) **Random Masking (RM)** (Thorne and Vlachos, 2021a): Randomly masks tokens without considering their contextual relevance. (b) **Heuristic Masking (HM)** (Thorne and Vlachos, 2021a): Masks only the tokens that are present in the claim but absent from the evidence (based on heuristics, these are more likely to be factually incorrect). (c) **Entity Masking (EM)** (Huang et al., 2023): It utilizes the entity set $E_s(\mathbf{x})$ (introduced in Section 3.1) to mask entities/phrases extracted from the input claim. For the claim mentioned in Figure 2, EM provides entities like ‘Pulmonary embolism’, ‘high blood oxygen levels’, ‘Pulmonary’ etc to mask. But this strategy does not consider redundancy or semantic importance of the phrases. Moreover, sometimes may mistakenly mask irrelevant tokens like “oxygen”, “indicated”, etc. Thus, relying only on entities may affect the model’s performance.

Our proposed **Diversity Masking (DM)**: It employs an MMR–based selection, which is a greedy algorithm and re-ranks entities in $E_s(\mathbf{x})$ to maximize their similarity to the claim while minimizing similarity between the selected tokens, thereby enhancing EM by ensuring both semantic relevance and diversity, as detailed in Section 3.1.

4.5 Methods Investigated

We compare our proposed methodology with the following baselines, considering two settings for each: 1) using evidence retrieved from an external corpus, and 2) using gold-standard evidence.

Non-parametric Baselines. These methods do not involve any parametric training for our downstream fact correction task. We employ the following: a) **0-shot** (Labruna et al., 2024; Kojima et al., 2022b; Li et al., 2023): To assess the importance of masking, in this approach, given a claim, the model directly prompts to generate the corrected claim without utilizing any example. b) **RAG** (Labruna et al., 2024): Similar to the previous approach, here we directly prompt the LLM to correct a given claim, but in the presence of some context. c) $\mathbf{M_2C_{w/Ver}}$: This variant incorporates an initial verification step in the M_2C framework by using an LLM, with corrections applied only to claims identified as incorrect. d) **ZeroFec-DA** (Huang et al., 2023), a training-free QA-based framework that divides the task into five sub-tasks. Here, RoBERTa (Liu et al., 2019) was fine-tuned on two biomedical datasets to improve domain-specific performance.

Distantly Supervised Baselines. We employ the following parametric approaches as baselines. a) **T5-Distant** (Thorne and Vlachos, 2021b): Similar to our methodology, it also follows a mask-then-correct approach using DPR (Karpukhin et al.,

			FEVER				SciFact			
			SARI (%)		BART		SARI (%)		BART	
LLM	Predictor	Ranker	Retrieved	Gold	Retrieved	Gold	Retrieved	Gold	Retrieved	Gold
n/a	ZeroFec-DA	MonoT5	39.3734	40.7713	-2.8629	-2.7623	32.1194	32.1988	-3.1905	-3.1756
	T5-Distant	GENRE	34.9734	43.7252	-5.4120	-2.4015	23.3640	22.9304	-3.0965	-2.9803
	CompEdit	MonoT5	26.0694	30.8499	-3.1801	-2.9100	30.4219	30.6196	-3.1699	-3.0591
Llama	0-shot	n/a	44.9646		-3.7971		35.0145		-2.9990	
	RAG	MonoT5	41.7222	42.7340	-2.3939	-2.0769	34.3641	36.5339	-2.6404	-2.3658
	$M_2C_{\text{w/Ver}}$	MonoT5	48.2327	49.5799	-2.6143	-2.4591	34.2985	30.6697	-2.9546	-3.1277
	M_2C_{DM}	BM25	45.0705	**51.2129**	-2.5935	-2.3720	34.8178	**40.0456**	-2.9862	**-2.6892**
		RM3	45.5321		-2.6088		36.8477		-2.9852	
		Contriever	48.6495		-2.6125		37.1026		-2.9409	
		ColBERT	47.4645		-2.6491		36.4724		-2.9212	
		MonoT5	49.1997		-2.6702		39.6940		-2.9304	
	$M_2C^{+}_{\text{DM}}$	n/a	49.3749		**-2.5939**		37.5820		-2.9302	
Qwen	0-shot	n/a	43.2337		-2.7315		35.7045		-3.0843	
	RAG	MonoT5	46.9218	48.5516	-2.7590	-2.5020	36.3252	34.7780	-2.4655	-2.8676
	$M_2C_{\text{w/Ver}}$	MonoT5	49.8118	51.0140	-3.0608	-2.4595	40.9380	42.6284	-2.7091	-2.6048
	M_2C_{DM}	BM25	45.9969	**51.2783**	-2.7278	**-2.3908**	35.6633	**42.6331**	-2.8656	**-2.1453**
		RM3	46.5141		-2.6143		39.1719		-2.9198	
		Contriever	49.2850		-2.6159		37.1282		-2.9282	
		ColBERT	48.3629		-2.6326		37.7490		-2.9198	
		MonoT5	50.3638		-2.6848		41.7093		-2.8450	
	$M_2C^{+}_{\text{DM}}$	n/a	50.7141		-2.5724		38.4647		-2.8636	

Table 2: Performance of M_2C^+ and M_2C in diversity-aware entity masking setting (i.e., 'DM' subscript) relative to the baselines. For each LLM, bold letters indicate the best performance among all models with retrieved/gold evidence, and the second-best results are underlined. In the table, all the RAG-based approaches (i.e., RAG, M_2C and M_2C^+) were obtained with a context size of 3, i.e., $p = 3$ in Equation 2. The results of M_2C^+ are in gray. Since MonoT5 performed best for M_2C setting, we chose MonoT5 for all the retrieval-based baseline experiments.

2020) and GENRE (De Cao et al., 2020) for evidence retrieval. It utilizes a heuristic masker and a finetuned T5-base corrector. b) **CompEdit:** (Fabbri et al., 2022): This method enhances summary factuality through post-editing. It uses BART model trained on sentence compression data to revise summaries by removing entities absent from the source document. These entities are detected using NER and highlighted with special tokens.

Our Proposed Approaches. We experiment with two variants of our framework: a) $\mathbf{M_2C}$, which employs individual rankers within the three-stage pipeline for selecting examples, and b) $\mathbf{M_2C^+}$, which ensembles corrected outputs from multiple rankers to mitigate noise and increase reliability.

4.6 Evaluation Metrics

We use BART score (Yuan et al., 2021) and SARI (Lewis et al., 2019) as our evaluation metrics. BART score measures semantic similarity between the generated prediction and the ground-truth evidence. SARI score helps to evaluate the quality of edits made by the corrector. In both cases, higher scores indicate more accurate and faithful corrections. A detailed description is provided in Appendix A. Additionally, to quantify the retrieval effectiveness of each IR model, we report nDCG@10 metric.

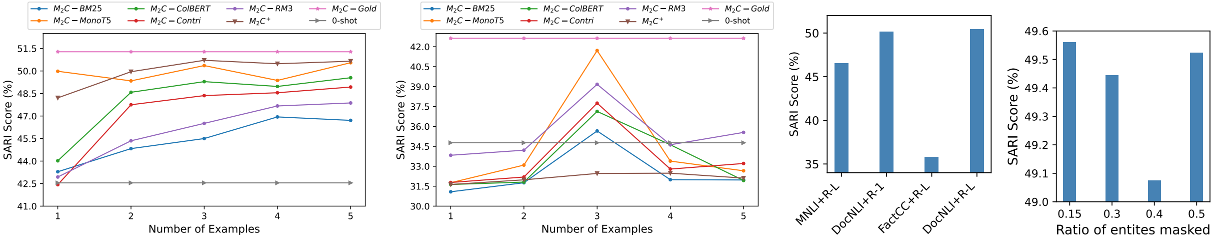
4.7 Implementation Details

We conduct all our experiments using the LLM Llama-2.0 (70B) (Touvron et al., 2023) and Qwen-2.5 (32B) (Yang et al., 2024), chosen after preliminary experiments with various open-source LLMs like LLaMA-3, Qwen-2, and Mistral etc. Notably, these models belong to two distinct decoder-only families, enabling evaluation across different architectures. For diversity-aware masking, we applied α as 0.3 and the number of masked sentences m as 10 (see Section 5.2) and used Pyserini (Lin et al., 2021) for retrieval. Further details have been provided in Appendix E. Our ensemble framework uses the four best-performing retrievers (excluding RM3) identified through preliminary analysis (discussed in Section 5.2).

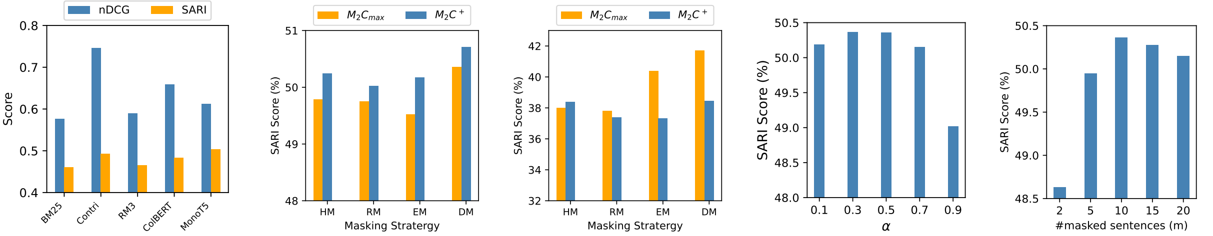
5 Result and Analysis

5.1 Main Observations

To address **RQ-1 (Performance of proposed methodology)**, Table 2 shows that both our proposed variants outperform all parametric and non-parametric baselines for each dataset, across both LLMs. Notably, our method outperforms the verification-based variant $M_2C_{\text{w/Ver}}$, indicating that the multi-stage correction pipeline already ensures strong factual alignment. On FEVER, $M_2C^{+}_{\text{DM}}$ performs better than M_2C_{DM} (shown in Figure 3f). In contrast, on SciFact, a dataset of scientific domain, M_2C_{DM} with MonoT5 performs better than



(a) Sensitivity on FEVER (b) Sensitivity on SciFact (c) Comparing Correction Scoring (d) Sensitivity of M_2C_{RM} on FEVER



(e) Correlation on FEVER (f) Comparison of four maskers on FEVER (g) Comparison of four maskers on SciFact (h) Sensitivity of α of M_2C_{DM} on FEVER (i) Sensitivity of m of M_2C_{DM} on FEVER

Figure 3: Sensitivity of M_2C variants using the Qwen model. (a–b) Sensitivity to the number of in-context examples on FEVER and SciFact; M_2C-Gold denotes use of gold-standard evidence. (c) Effect of different correction scoring combinations on M_2C_{DM}–MonoT5 for FEVER. (d) Sensitivity of token masking ratio of M_2C_{RM}–MonoT5 on FEVER. (e) Correlation between retrieval quality (nDCG@10) and downstream performance (SARI(%), normalized by dividing them by 100) for M_2C_{DM}. (f–g) Comparison of masking strategies for M_2C variants on FEVER and SciFact; M_2C_{max} indicates the best performance of M_2C across all retrievers. (h–i) Effect of diversity parameter α and number of masked sentences (m) in M_2C_{DM}–MonoT5 on FEVER.

$M_2C^+_{DM}$ (shown in Figure 3g). The likely reason is that in specialized domains, a single high-performing retriever (like MonoT5, in Table 2) may already retrieve sufficiently relevant evidence, whereas ensembling may lead to the selection of non-relevant evidence retrieved by multiple under-performing retrievers, thereby degrading the performance. Moreover, the effectiveness of our model in real-world scenarios is discussed in Appendix D.

In relation to **RQ-2 (Choice of masker)**, in this work, we experiment with four masking strategies: Random (RM), Heuristic (HM), Entity (EM), and our proposed Diversity-aware Masking (DM). Figures 3f and 3g compare the performance (SARI score (%)) of M_2C_{max} (reports the best performance of M_2C across five retrievers) and M_2C^+ models across these masking strategies. We observe that, for both datasets, DM and HM generally outperform RM and EM in both M_2C variants, although DM, EM outperform HM and RM on the SciFact dataset under the M_2C_{max} setting. In FEVER, which is a general domain dataset, DM significantly outperforms HM for both M_2C variants. The likely reason is that MMR effectively balances semantic relevance and diversity, allowing it to handle noisy or paraphrased evidence. In contrast, HM strictly relies on token-level overlaps, which leads to false negatives. On the other hand, in the SciFact dataset, collected from the scientific domain, HM's direct token matching is more effective in scientific texts. Moreover, the limited linguistic variability in SciFact reduces the benefit of selecting diversity-based entities. Thus, DM achieves comparable result to HM for both models.

In relation to **RQ-3 (Gold vs. retrieved evidence)**, all models perform better with gold evidence, which serves as an oracle setup due to its higher relevance. However, there is less difference in performance. Specifically, our proposed M_2C variants, M_2C_{DM} and $M_2C^+_{DM}$, achieve comparable results with both gold and retrieved evidence, demonstrating the effectiveness of our method without relying on annotated evidence. Notably, for M_2C_{DM} and $M_2C^+_{DM}$, the results are the same with gold evidence.

To address **RQ-4 (Correlation between retriever and performance)**, Figure 3e demonstrates a positive correlation between retrieval quality (nDCG@10) and downstream correction performance (SARI) for each retriever, indicating that higher-quality retrievals lead to more accurate corrections. As shown in Table 3, retriever choice substantially affects correction quality–retrievers like MonoT5 and ColBERT retrieve the most relevant evidence for given claims, while Contriever achieves the highest overall nDCG@10. Therefore, retrievers vary in effectiveness depending on the topic or the claim type. Thus, motivated by these

Retrieved evidences	nDCG	Time
Input Claim: One Dance was by a Mexican. **Ground Truth:** One Dance was by a Canadian.		
BM25: The Jarabe Tapatío, better known internationally as "The Mexican Hat Dance" ...	0.5767	20–25
Contriever: Aspen Santa Fe Ballet ASFB is an American contemporary dance company...	**0.7465**	270
RM3: One Dance is a 2003 Canadian romantic drama film about ...	0.5892	20–30
ColBERT: "One Dance" is a song by Canadian rapper Drake from his fourth studio album...	0.6590	192
MonoT5: "One Dance" is a song by Drake, a Canadian rapper, singer, and songwriter...	0.6122	250

Table 3: Examples of top-most similar sentences retrieved using different rankers from the FEVER dataset. These evidence are used in prompts for correcting the input claims. Also, we report nDCG@10. The blue highlighted parts denote correctly retrieved phrases and red parts denote the incorrect phrases responsible for the erroneous prospective corrections. Reported retrieval times (in seconds) correspond to 100 claims.

		Time				
	Memory	Ranking		Model		Total
Model	(max)	Tr	Infer	Tr	Infer	
ZeroFec-DA	8GB	-	2.70	-	6.5	6.5
T5-Distant	48GB	18.0	2.70	2.10	1.00	23.80
CompEdit	5GB	-	2.70	18.50	1.00	23.20
M_2C_{DM}	48GB	-	2.70	-	7.5	10.20
$M_2C^+_{DM}$	48GB	-	7.52	-	30.00	37.52

Table 4: GPU memory and total time (in hours) for processing all claims on FEVER dataset with Qwen as base model.

observations, we proposed M_2C^+, an ensemble-based strategy that aggregates evidence from multiple retrievers to harness their complementary strengths and reduce noise of individual retrievers.

5.2 Sensitivity Analysis

Sensitivity of the Number of Examples. In FEVER dataset, we observe from Figure 3a, M_2C^+ exhibits greater stability with respect to context size p (see Equation 2) (i.e., the number of retrieved examples), consistently outperforming all individual retriever variants of M_2C and simple 0-shot. In contrast, for the SciFact dataset, M_2C^+ underperforms compared to M_2C with RM3 and MonoT5 as context size increases, likely due to domain-specific vocabulary and token mismatches affecting retrieval relevance (see Figure 3b). Also for SciFact, across most experiments, the optimal performance is observed for $p = 3$, which we have reported in Table 2. However, for both datasets, M_2C with gold annotated evidence, i.e., M_2C-Gold, being an oracle setup, achieves the highest performance.

Method	BM25	RM3	Contriever	ColBERT	MonoT5	SARI	BART
M_2C^+	✓	✗	✗	✓	✓	50.4832	-2.5809
	✓	✗	✓	✗	✓	50.4231	-2.5774
	✓	✗	✓	✓	✗	49.7552	-2.6009
	✓	✓	✓	✓	✗	49.7772	-2.5925
	✓	✓	✓	✗	✓	50.0614	-2.5852
	✓	✓	✗	✓	✓	50.0763	-2.5907
	✓	✗	✓	✓	✓	**50.7141**	-2.5724
	✗	✓	✓	✓	✓	50.6295	**-2.5701**
	✓	✓	✓	✓	✓	50.4485	-2.5779

Table 5: Retriever sensitivity analysis for M_2C^+ on the FEVER dataset using Qwen as the base model.

Case Study on Scoring Function. Figure 3c represents an ablation of the impact of different scoring configurations on the FEVER test set with a balancing factor of $\lambda = 0.5$ (see Appendix F). Inspired by Manakul et al. (2023), we use DeBERTa-v3-large (He et al., 2021) fine-tuned to MNLI as an entailment model with the ROUGE-L metric. Following the metric combinations used by Huang et al. (2023), we have compared the performance of the correction scoring in Figure 3c. We observe that combining DocNLI (Yin et al., 2021) and ROUGE-L performs best in terms of SARI and BART score.

Computational Analysis. Table 4 shows GPU memory and total processing time for all claims on FEVER. We observe that although, $M_2C_{DM}{}^+$ incurs a moderate latency increase (2.7–4×) than M_2C_{DM} due to multi-retriever aggregation, it provides only marginal gains over M_2C_{DM}, indicating that the faster variant remains highly effective. The retrieval component is CPU-based, and the indexing is performed once per corpus, minimizing recurring computational overhead. Both are inference-only and maintain GPU usage (48GB), avoiding the heavy training overhead like baselines. Overall, our framework provided an efficient balance between computational cost and correction effectiveness.

Sensitivity of Masking Strategies. We analyze key hyperparameters for both masking strategies: the token masking ratio in RM, and the relevance–diversity trade-off (α, Eq. 1) along with the number of selected entities (m) in DM. For RM, masking 15% of tokens yields the best performance (Figure 3d), consistent with BERT (Devlin et al., 2019), while higher ratios degrade performance due to information loss. For DM, $\alpha = 0.3$ provides the optimal trade-off between relevance and diversity (Figure 3h), while number of masked

Example 1 (MAJORITY VOTING FAILURE)	
Input: "There are currently 417 Mormon members as of 2017."	*Ground Truth:* "There are currently 15,882,417 Mormon members as of 2017."
One of the Candidate Answers: "There are currently 15,882,417 Mormon members as of 2017."	*Final Answer:* "There are 64,123 members as of 2017."
Note: The correct candidate is present but outvoted.	
Example 2 (ENTITY MISMATCH IN CORRECTION)	
Input: "LinkedIn is based in Spain."	*Ground Truth:* "LinkedIn is based in the United States." *Final Answer:* "LinkedIn is not based in Spain."
Note: Negation is correct, but it fails to provide the correct entity.	
Example 3 (CORRECTION SCORING FAILURE)	
Input: "The Giver is a TV show."	*Ground Truth:* "The Giver is a film."
Candidate Correction 1: "The Giver is a TV show."	DocNLI: 0.975, ROUGE-L: 1.0
Candidate Correction 2: "The Giver is a film."	DocNLI: 0.997, ROUGE-L: 0.767
	Final Answer: "The Giver is a TV show."
Note: Correct option has higher factual score but loses due to lower ROUGE.	

Table 6: Representative error cases in M$_2$C$^+$: failures due to majority voting, entity mismatch, and limitations of the correction scoring scheme.

sentences in M$_2$C$_{\mathrm{DM}}$, $m = 10$ achieves the best balance between informativeness and noise reduction (Figure 3i). We use these setting for all the subsequent experiments.

Sensitivity of retrievers in M$_2$C$^+$. Table 5 reports the sensitivity of M$_2$C$^+$ to different retriever combinations on the FEVER using Qwen as the base model. The results show that dense retrievers such as Contriever, ColBERT, and MonoT5 consistently strengthen the ensemble by retrieving semantically similar evidence, particularly under low lexical overlap. But, the inclusion of RM3 often underperforms due to its reliance on pseudo-relevance feedback. Also, combinations excluding RM3 but incorporating at least two dense retrievers provide the best results. The best SARI score is achieved when combining all retrievers except RM3, while excluding BM25 yields the best BART score. All the subsequent experiments are reported without RM3. As M$_2$C$^+$ is a majority voting based method, at least three retrievers are required to reach a decision; hence, we progressively constructed an ensemble starting with any three retrievers, then four, and finally all five. Thus, ensembling diverse retrievers effectively mitigates individual retriever biases and enhances downstream performance.

5.3 Error Analysis

Table 6 shows some examples where our model failed. Example 1 shows the limitation of the majority voting approach, i.e., M$_2$C$^+$. Here, the desired correction was generated by a few retrievers but was ultimately discarded due to being outvoted by the erroneous prospective corrections from others. In example 2, the model correctly identifies that the claim is wrong (e.g., LinkedIn is not based in Spain) but fails to replace the incorrect part with the correct information (i.e., 'the United States'). Although our model produced a logically valid correction, the model deviated from the desired output. Example 3 shows the failure of the correction scoring function (see Section 3.3). The model chooses a less factual candidate because it has a higher ROUGE score, even though the correct candidate has more factual alignment. By addressing the genesis of these errors we can make the factual correction model more reliable.

Some errors arise from annotation inconsistencies in the benchmark datasets, causing the evaluation to unfairly penalize our model's outputs; we discuss these cases in detail in Appendix C.

6 Conclusions and Future Work

We propose Mask-to-Correct (M$_2$C), a training and annotation-free RAG-based framework for faithful fact correction. Here, we introduce a diversity-aware masking strategy for selecting diversified but relevant entities. Furthermore, we propose an ensemble approach, M$_2$C$^+$, that aggregates corrections across diverse retrievers. Extensive experiments on FEVER and SciFact datasets demonstrate that our method consistently outperforms all baselines. In future, to improve transparency and trustworthiness, we plan to extend M$_2$C to a self-correcting framework by incorporating a feedback loop, where generated corrections will guide retriever reranking and masking.

Acknowledgment. We used Gen AI tools for minor paraphrasing. Final content was reviewed.

Limitation

A limitation is that the computation is memory-intensive as one needs to load the LLMs, thus making it difficult to execute our approach on a terminal with limited memory capacity. Additionally, due to the limitations of our own resources, we were not successful in employing a larger or a commercially accessible LLM (e.g., GPT-4, Claude).

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. *arXiv preprint arXiv:2204.10290*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46.

Dhananjay Ashok, Atharva Kulkarni, Hai Pham, and Barnabás Póczos. 2023. Scifix: Outperforming gpt3 on scientific factual error correction. *arXiv preprint arXiv:2305.14707*.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyi, Samira Khorshidi, Fei Wu, Ihab F Ilyas, and Yunyao Li. 2023. Fleek: Factual error detection and correction with evidence retrieved from external knowledge. *arXiv preprint arXiv:2310.17119*.

Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. *arXiv preprint arXiv:1801.04470*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Jiangjie Chen, Rui Xu, Wenxuan Zeng, Changzhi Sun, Lei Li, and Yanghua Xiao. 2023. Converge to the truth: Factual error correction via iterative constrained editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12616–12625.

Anton Chernyavskiy and Dmitry Ilvovsky. 2019. Extract and aggregate: A novel domain-independent approach to factual data verification. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 69–78, Hong Kong, China. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexander R Fabbri, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu, and Caiming Xiong. 2022. Improving factual consistency in summarization with compression-based post-editing. *arXiv preprint arXiv:2211.06196*.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Ulrich Germann. 2003. Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 72–79.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300*.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Xiao Han, Yuqi Liu, and Jimmy Lin. 2021. The simplest thing that can possibly work: (pseudo-)relevance feedback via text classification. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, page 123–129, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Xingwei He, Qianru Zhang, A-Long Jin, Jun Ma, Yuan Yuan, and Siu Ming Yiu. 2024. Improving factual error correction by learning to inject factual errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18197–18205.

Christopher Hidey and Mona Diab. 2018. Team sweeper: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 150–155.

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023. Zero-shot faithful factual error correction. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. 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 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Volume 2: Short Papers)*, pages 402–410.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022a. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022b. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Tiziano Labruna, Jon Ander Campos, and Gorka Azkune. 2024. When to retrieve: Teaching llms to utilize information retrieval effectively. *arXiv preprint arXiv:2404.19705*.

Tian Lan, Deng Cai, Yan Wang, Heyan Huang, and Xian-Ling Mao. 2023. Copy is all you need. *ICLR*, abs/2307.06962.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Mingda Li, Xinyu Li, Yifan Chen, Wenfeng Xuan, and Weinan Zhang. 2024. Unraveling and mitigating retriever inconsistencies in retrieval-augmented large language models. *arXiv preprint arXiv:2405.20680*.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023. From classification to generation: Insights into crosslingual retrieval augmented icl. *arXiv preprint arXiv:2311.06595*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *arXiv preprint arXiv:2102.10073*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Andrew Parry, Sean MacAvaney, and Debasis Ganguly. 2024. Exploiting positional bias for query-agnostic generative content in search. *CoRR*, abs/2405.00469.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Payel Santra, Madhusudan Ghosh, Debasis Ganguly, Partha Basuchowdhuri, and Sudip Kumar Naskar. 2024. “the absence of evidence is not the evidence of absence”: Fact verification via information retrieval-based in-context learning. In *Big Data Analytics and Knowledge Discovery: 26th International Conference, DaWaK 2024, Naples, Italy, August 26–28, 2024, Proceedings*, page 381–387, Berlin, Heidelberg. Springer-Verlag.

Payel Santra, Madhusudan Ghosh, Debasis Ganguly, Partha Basuchowdhuri, and Sudip Kumar Naskar. 2025. The “curious case of contexts” in retrieval-augmented generation with a combination of labeled and unlabeled data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2):e70021.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Darsh Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8791–8798.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 359–366. Springer.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on FEVER*, pages 105–109, Hong Kong, China. ACL.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

James Thorne and Andreas Vlachos. 2021a. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, Online. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2021b. Evidence-based factual error correction. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pages 3298–3309. Association for Computational Linguistics (ACL).

James Thorne and Andreas Vlachos. 2021c. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2022. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. *arXiv preprint arXiv:2406.09815*.

Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023. Repocoder: Repository-level code completion through iterative retrieval and generation. *arXiv preprint arXiv:2303.12570*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.

A Metric Descriptions

A.1 SARI Score

SARI (Xu et al., 2016) is designed for tasks where a system must edit an input sentence (e.g., factual correction), rather than generate text from scratch. Unlike overlap-based metrics such as BLEU or ROUGE, SARI explicitly evaluates the goodness of the edits by comparing the input sentence, system output (predicted correction), and reference (i.e., ground truth or retrieved supporting context).

SARI is computed using three components as,

$$\text{SARI} = \frac{1}{3}(\text{F}_{\text{add}} + \text{F}_{\text{keep}} + \text{P}_{\text{del}})$$

where, Fadd assesses the addition of relevant n-grams present in the reference but absent in the input, Fkeep evaluates the preservation of correct input n-grams, and P_{del} measures the removal of incorrect or unnecessary content. In our experiments, we use $n = 4$ for n-gram computation.

By utilizing add, keep, and delete operations, SARI rewards only edits that align with ground-truth corrections and penalizes incorrect or excessive rewriting, making it well-suited for factual correction tasks where minimal yet accurate changes are desired.

A.2 BARTScore

BARTScore (Yuan et al., 2021) is a reference-based evaluation metric that measures the relationship between two text sequences using a pretrained sequence-to-sequence model. It computes the log-likelihood of generating a target sequence conditioned on a source sequence, thereby capturing how well one text is supported by another.

Formally, given an input claim $\mathbf{x}$ and a reference evidence sentence e, the BARTScore is defined as:

$$\text{BARTScore}(\mathbf{x}, e) = \frac{1}{|e|} \sum_{t=1}^{|e|} \log P(e_t \mid e_{<t}, \mathbf{x}),$$

where $P(\cdot)$ is computed using a pretrained BART model.

In our work, we use BARTScore to measure how well our correction is supported by the gold evidence. Specifically, the score reflects the extent to which the corrected claim can account for or generate the supporting evidence, thereby indicating its factual grounding.

Annotation Error

Example 1

Input Claim: Exit the King is by a man.
Ground Truth: Exit the King is by a man.

Our Correction: Exit the King is by Eugène Ionesco.
Supporting Evidence: Exit the King is a play by Eugène Ionesco, originally written in French in 1962.

Example 2

Input Claim: A U.S. state is Phoenix, Arizona.
Ground Truth: Phoenix, Arizona is a U.S. state.

Our Correction: A U.S. state is Arizona.
Supporting Evidence: Arizona is a state located in the southwestern region of the United States.

Table 7: Example of an annotation error in the dataset. The blue-highlighted spans indicate correct factual information, while the red highlights indicate incorrectly accepted text in the original claim and annotation.

A.3 DocNLI

DocNLI (Yin et al., 2021) is a document-level natural language inference task in which a model determines whether a claim $\mathbf{x}$ is entailed by, contradicted by, or unsupported with respect to an evidence document $e \in \mathcal{E}(\mathbf{x})$. Formally, given a pair $(\mathbf{x}, e)$, the classifier outputs one of three labels: $y \in \{\text{entailment}, \text{contradiction}, \text{neutral}\}$.

In our work, we use a RoBERTa [3] model fine-tuned on the DocNLI dataset and use the entailment probability, i.e.,

$$\text{Entail}(\mathbf{x}, e) = P(y = \text{entailment} | \mathbf{x}, e),$$

as a factuality score for candidate corrections. In this paper, this score captures the degree of factual alignment between the corrected claim and the retrieved evidence, and is used as a component of

---

[3] `https://huggingface.co/FacebookAI/roberta-large`

Input Claim	**GT Claim (TRUE)**	**Model Output**	**Observation**	**SARI**
Jenny McCarthy modeled for Playboy photographers.	Same as input	Jenny McCarthy modeled for Playboy magazine.	Semantics preserved; minor paraphrase penalized.	28.79
An American directed One True Thing.	Same as input	Carl Franklin directed One True Thing.	Output is more specific and factually correct; SARI penalizes ADD.	22.64
The left hand side of the Trebbia River…	Same as input	The left bank of the Trebbia River…	Bank vs side; paraphrased → SARI penalizes.	28.82
Great white sharks have never killed people.	Same as input	Great white sharks have killed people.	GT claim is factually incorrect, so SARI unfairly penalizes correction.	22.45
Exit the King is by man.	Same as input	Exit the King is by Eugène Ionesco.	GT claim is incorrect and this claim is ambiguous.	28.29

Table 8: Examples demonstrating the limitations of SARI on TRUE-labelled instances (input equals ground truth). The metric penalizes factually correct outputs due to its sensitivity to surface-form changes, including valid paraphrases, informative refinements, and inconsistencies or errors in the ground-truth annotations.

	SARI (%)		BART		Overall	
Model	TRUE	FALSE	TRUE	FALSE	SARI	BART
M_2C	31.5323	63.4833	-2.6904	**-2.4995**	50.3638	-2.6848
M_2C^+	**34.1467**	**64.0014**	**-2.6905**	-2.4900	**50.7141**	**-2.5724**

Table 9: Class-wise performance of M_2C (MonoT5 as retriever) and M_2C^+ on the FEVER dataset with Qwen as base LLM. Bold values indicate the best performance between M_2C and M_2C^+.

the correction scoring function to rank candidate outputs based on their entailment likelihood (as described in Section 3.3).

A.4 ROUGE-L

ROUGE-L (Lin, 2004) measures the similarity between a generated sentence and a reference based on the length of their Longest Common Subsequence (LCS), capturing both lexical overlap and word order.

Formally, given a candidate correction $\mathbf{x}'$ and the original claim $\mathbf{x}$, the ROUGE-L score is computed using LCS-based precision and recall, which are combined into an F1-score.

In our work, we use ROUGE-L as a proxy for edit minimality, i.e., to quantify how much the corrected candidate deviates from the input claim. Similar to *ZeroFEC* (Huang et al., 2023), we use this score in conjunction with the DocNLI-based entailment score to rank candidate corrections.

Specifically, while the DocNLI score captures factual consistency with respect to external evidence, ROUGE-L encourages minimal and conservative edits by favoring candidates that preserve the structure and content of the original claim. The combination of these two signals ensures that the selected correction is both factually grounded and minimally deviating from the input, thereby maintaining faithfulness while correcting errors.

B Dataset Description

In this paper, we conduct experiments on two benchmark datasets: FEVER (Thorne and Vlachos, 2021c) and SciFact (Wadden et al., 2020) datasets. We use these fact correction datasets introduced by (Huang et al., 2023), where the supported claims are taken as faithful, and unfaithful versions are generated by applying Knowledge Base Informed Negations (Wright et al., 2022) to a subset of the faithful claims. Moreover, we used the FEVER validation data (Thorne and Vlachos, 2021a) to conduct an **extensive grid search** to optimize the hyperparameters for our downstream fact correction task.

C Information of our Annotated data

Table 7 illustrates examples of annotation errors in the benchmark dataset. In Example 1, the input claim “Exit the King is by a man” is marked as correct, although the accurate correction should be Eugène Ionesco; here, “man” is a generalized term. Similarly, in Example 2, the claim “A U.S. state is Phoenix, Arizona” is truth claimed by annotators, but originally Phoenix is a city in the state of Arizona. In both case, our model correctly generates predicted outputs using retrieved evidence, but is penalized due to wrong gold annotations.

D Real-World Applicability

FEVER and SciFact include both true and false as well as ambiguous claims resembling real-world scenarios. For instance, examples such as “Parkinson’s disease causes symptoms", “Shut Up is a title", “Benzodiazepines can be taken", and “Fraud can be used for gain" are ambiguous, context-dependent claims and require the retrieval of appropriate evidence for correction. In practical settings,

Prompt Template

```
Given a claim and related evidence, your task is to correct it if the claim is not supported by
the given evidence. If the input claim is correct, do not edit it and give the input claim as
output. Your output claim should be faithful to the evidence and should not deviate much from
the input claim. Do not print anything else in the output except the corrected claim. Strictly
follow the syntax given below for output syntax:
Input Claim: [sentence]
Evidence: [document]
Output Correction: [sentence]
```

Selected Examples

```
[1.] Pulmonary embolism (PE) is a blockage of an artery in the lungs by... Signs of a PE include
low blood oxygen levels.
[2.] ...
```

Test Instance

```
Input Claim: Pulmonary embolism is indicated by high blood oxygen levels.
Output Correction:
```

Figure 4: An illustration of the prompt structure used in the few-shot RAG experiment.

a fact-correction system must not only revise false claims but also preserve correct ones and appropriately handle ambiguous cases. To evaluate this behavior, we analyze the effectiveness of our framework across different classes, as shown in Table 9.

We observe that, M$_2$C and M$_2$C$^+$ effectively preserve correctness for true claims while correcting false ones, thus the framework does not operate under the assumption that all inputs are incorrect and hence generalizing well in real-world scenarios.

Moreover, we observe that SARI$(True)$ $<$ SARI$(False)$. This arises from the nature of SARI as an edit-based metric, which rewards necessary modifications but penalizes unnecessary changes. Since FALSE claims require edits, they achieve higher scores, whereas TRUE claims ideally remain unchanged, leading to penalization even for minor paraphrases.

As shown in Table 8, our manual inspection reveals that the observed performance degradation is sometimes due to annotation inconsistencies and underspecified ground-truth claims, which cause SARI to penalize factually correct outputs.

E Implementation Details

We conduct all our experiments using the LLM Llama-2.0[4] (70B) (Touvron et al., 2023) and Qwen-2.5[5] (32B) (Yang et al., 2024) model. For our proposed diversity-aware masking-based methods, like M$_2$C$_{\text{DM}}$ and M$_2$C$^+_{\text{DM}}$, we used all-MiniLM-L6-v2 for getting embeddings of entities and used KeyBERT[6] (Grootendorst, 2020; Bennani-Smires et al., 2018). For entity-masker, we extract the named entities using SpaCy[7] (Vasiliev, 2020) and verb, adjective, and noun phrases using Stanza[8] (Qi et al., 2020), and subsequently mask them. Next, for all our LLM-based M$_2$C and M$_2$C$^+$ experiments, we utilize the vLLM (Kwon et al., 2023; Lin et al., 2024) library to apply k-v cache optimization, enhancing computation speed. For fine-tuning the supervised baselines in our experiments (namely, T5-distant, CompEdit, and ZeroFec-DA), we follow the respective setups as reported in their original works. T5-distant uses a heuristic masking strategy and is fine-tuned on randomly masked data for 10 epochs with a learning rate of $5e-5$. CompEdit employs the publicly available BART-large (Lewis et al., 2019) post-editor trained for 10 epochs with batch size 64 on compression data generated using a BART-based perturber, and inference is performed using greedy decoding (Germann, 2003). In our implementation, we used SpaCy (Vasiliev, 2020) and Stanza (Qi et al., 2020) for entity recognition. For ZeroFec-DA, we use the domain adaptation variant, where the DocNLI model is fine-tuned on PubMedQA (Jin et al., 2019) and BioASQ (Tsatsaronis et al., 2015) datasets for up to 5,000 steps using AdamW (Loshchilov and Hutter, 2019) with a learning rate of $3e-6$. All generative models use beam search (Freitag and Al-Onaizan, 2017) with a beam width of 4 during inference. All experiments throughout the paper were performed using

---

[4] https://huggingface.co/TheBloke/Llama-2-70B-Chat-AWQ

[5] https://huggingface.co/Qwen/Qwen2.5-32B-Instruct-AWQ

[6] https://github.com/MaartenGr/KeyBERT

[7] https://spacy.io/

[8] https://stanfordnlp.github.io/stanza/

a NVIDIA A6000 (48GB) GPU.

	λ				
Metric	0.0	0.2	0.5	0.8	1.0
SARI (%)	21.0695	40.7768	**52.7064**	51.7891	50.5760

Table 10: Ablation study on the effect of the balancing factor λ in our M_2C framework using MonoT5 ranker on the FEVER validation set with Qwen model. The table reports SARI scores (%) for different values of λ.

F Ablation on the balancing factor λ

We study the impact of the balancing factor λ, which controls the trade-off between factuality and faithfulness. As shown in Table 10, $\lambda = 0.5$ gives the best performance in terms of SARI score for M_2C-MonoT5 experiment. Thus we adopt this value for all experiments throughout the paper.

G Prompt Templates

Figure 4 shows the template used for the 0-shot and RAG experiment, and Figure 5 shows the templates we used in the evidence-guided corrector part of M_2C framework.

Prompt Template

```
Your task is to correct a claim by filling in the [MASK] using the provided input evidence,
ensuring that the corrected claim is supported by the evidence and only differs from the input
claim in the masked positions. If the input claim is correct, do not edit it and give the input
claim as output. Your output claim should be faithful to the provided evidence and should not
deviate much from the input claim.
Please use the most relevant evidence to correct the claim.  The corrected claim shouldn't
contain any [MASK].

Input Claim: [sentence]
Evidence: [document]
Masked Claim: [masked sentence]
Output Correction: [sentence]
```

Selected Examples

```
[1.] Pulmonary embolism (PE) is a blockage of an artery in the lungs by...Signs of a PE include
low blood oxygen levels.
[2.] ...
```

Test Instance

```
Input Claim: Pulmonary embolism is indicated by high blood oxygen levels.
Masked Claim: Pulmonary embolism is indicated by [MASK].
Output Correction:
```

Figure 5: An illustration of the prompt structure used in our proposed approach M_2C.