Article

Context-Grounded Factuality Enhancement in LLM Responses via Multi-Stage Critique and Refinement

Salma Ali *

Posted Date: 12 August 2025

doi: 10.20944/preprints202508.0700.v1

Keywords: large language models; chain-of-thought reasoning; self-correction; zero-shot prompting; natural language processing

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 August 2025 doi:10.20944/preprints202508.0700.v1

Article

Context-Grounded Factuality Enhancement in LLM Responses via Multi-Stage Critique and Refinement

Salma Ali

Universiti Teknologi Malaysia; hva180033@siswa365.um.edu.my

Abstract

Large Language Models (LLMs) often suffer from factual hallucinations and contextual detachment, significantly limiting their reliability in critical applications. To address these issues, we propose an innovative automated framework, "Context-Grounded Factuality Enhancement in LLM Responses via Multi-Stage Critique and Refinement." Our method leverages the inherent reasoning capabilities of pre-trained LLMs themselves, operating in a zero-shot manner without requiring any fine-tuning. It simulates a "Fact Verifier-Content Reviser" role within the LLM, guiding it through a multi-stage Chain-of-Thought (CoT) reasoning process to systematically identify, classify, and correct factual inconsistencies and ungrounded statements against provided source documents. Evaluated on challenging datasets, HotpotQA and ELI5, our framework significantly outperforms baseline LLMs and existing simple self-correction strategies in terms of Fact Consistency Score (FCS) and Context Grounding Score (CGS). Notably, our CoT-guided prompting strategy consistently yields superior results, achieving state-of-the-art performance with Llama 3 70B. Human evaluations further corroborate the enhanced factual accuracy and contextual grounding, alongside maintained fluency. While involving increased computational cost due to explicit reasoning, our framework demonstrates a robust and effective approach to improving the trustworthiness of LLM-generated content.

Keywords: large language models; chain-of-thought reasoning; self-correction; zero-shot prompting; natural language processing

---

I. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating coherent and fluent text, revolutionizing various applications from content creation and conversational AI to specialized domains like multi-modal content generation [1] and industrial defect recognition [2]. However, a significant challenge persists: their propensity for *factual hallucinations* and *contextual detachment* [3]. Despite their impressive fluency, LLMs frequently "fabricate" information or produce responses that are inconsistent with provided contexts or external knowledge sources, often without explicit instruction. This unreliability severely limits their applicability in high-stakes domains such as healthcare [4], legal services, and technical support, where factual accuracy and trustworthiness are paramount. The ability to generate factually consistent and contextually grounded responses is thus critical for the widespread adoption and responsible deployment of LLMs.

Addressing these critical limitations, our work introduces an innovative automated framework, titled **"Context-Grounded Factuality Enhancement in LLM Responses via Multi-Stage Critique and Refinement"**. We are motivated by the need to develop robust mechanisms that enable LLMs to automatically identify and correct their own factual errors and ungrounded statements, thereby enhancing their overall reliability. Current approaches often involve fine-tuning or external knowledge retrieval, which can be resource-intensive or limited by the scope of the knowledge base. Our approach distinguishes itself by leveraging the inherent reasoning capabilities of pre-trained LLMs themselves to perform self-critique and revision in a zero-shot manner.

Our proposed method, referred to as **Ours**, builds upon the principle of self-correction by simulating a "Fact Verifier-Content Reviser" role within the LLM architecture. This is achieved through a multi-stage Chain-of-Thought (CoT) [5] reasoning process, which guides the LLM to systematically analyze its initial output, or that of another LLM, against provided factual sources and internal logical consistency. Specifically, the framework prompts the LLM to first identify potential inconsistencies or hallucinations, then pinpoint the exact erroneous segments, classify the error type (e.g., factual inaccuracy, ungrounded statement), and finally, propose a revised response that aligns with the factual evidence. A key advantage of our methodology is its *zero-shot prompting* nature, which means no additional training or fine-tuning of the LLM is required. We rely solely on the pre-trained LLM's generalization abilities and instruction-following capabilities, driven by meticulously designed prompts.

To thoroughly evaluate the effectiveness of our framework in enhancing factual consistency and contextual grounding, we conducted experiments on two challenging and widely recognized datasets:

- **HotpotQA** [6]: A multi-hop question answering dataset that necessitates aggregating information from multiple documents to answer questions. This dataset is particularly suitable for assessing an LLM's ability to process complex information and anchor its responses to diverse contexts.
- **ELI5 (Explain Like I'm 5)** [7]: A dataset designed for generating simplified explanations of complex concepts. We utilized its longer answer segments to evaluate the model's capacity to maintain explanatory accuracy without introducing spurious information.

Our evaluation focused on two key metrics: Fact Consistency Score (FCS) and Context Grounding Score (CGS), both normalized between 0 and 1, with higher scores indicating better performance. We compared our method against baseline LLMs (without correction) and existing simple self-correction strategies. The experimental results demonstrate the superior performance of our approach. Notably, our method, particularly when employing more complex CoT prompting strategies, consistently achieved state-of-the-art results on both HotpotQA and ELI5 datasets, significantly outperforming all baselines and competitive models in terms of both factual consistency and contextual grounding. This underscores the efficacy of our multi-stage critique and refinement framework, especially when combined with powerful LLMs and explicit CoT reasoning.

In summary, our main contributions are as follows:

- We propose a novel multi-stage critique and refinement framework that enables LLMs to automatically enhance the factual consistency and contextual grounding of their generated responses.
- We demonstrate the effectiveness of leveraging LLMs themselves as sophisticated "Fact Verifier-Content Reviser" agents through zero-shot Chain-of-Thought prompting, eliminating the need for additional model training.
- Our method achieves state-of-the-art performance in improving factual consistency and context grounding on challenging question answering and explanation generation benchmarks, without requiring any fine-tuning of the underlying LLMs.

II. Related Work

A. Enhancing Factual Consistency and Grounding in Large Language Models

This subsection reviews recent advancements in enhancing the factual consistency and grounding of Large Language Models (LLMs). A comprehensive survey by Lei et al. [8] provides a novel taxonomy of hallucination in LLMs, identifying contributing factors and reviewing detection and mitigation strategies, which is directly relevant to understanding and addressing factual hallucinations, including those pertinent to retrieval-augmented systems and knowledge boundaries. Building on this, Wrick et al. [9] introduce a novel framework for achieving **contextual grounding** in LLMs, focusing on explicit context representation using knowledge representation and reasoning techniques to enhance reliability and ethical alignment, aiming to improve LLM fidelity by anchoring model behavior within situational, cultural, and ethical contexts. Such efforts are bolstered by advancements in pre-trained models for event correlation and context-to-event generation, which enhance the understanding and generation of structured information [10,11]. Furthermore, Isabelle et al. [12] survey how generative LLMs can support human fact-checkers, exploring various prompting and fine-tuning techniques for fact verification tasks and highlighting existing methods and their limitations. In a similar vein, Yizheng et al. [13] provide a comprehensive survey of retrieval-augmented text generation techniques for LLMs, detailing advancements in using LLMs for text-to-SQL generation and categorizing methods based on training strategies. To address factual consistency and grounding, Linyiao et al. [14] propose Knowledge Graph-Enhanced Large Language Models (KGLLMs) to integrate explicit factual knowledge from Knowledge Graphs (KGs) into LLMs, thereby improving their factual reasoning capabilities and developing models more adept at fact-aware content generation. Complementing this, other works have explored modeling event-pair relations in external knowledge graphs for enhanced script reasoning [15]. Bishwamittra et al. [16] further investigate the logical consistency of LLMs specifically within the context of fact-checking, exploring their utility in supporting human fact-checkers to improve information consistency. Moreover, Ehsan et al. [17] survey existing Open-Domain Question Answering benchmarks, offering a taxonomy of datasets and a structured analysis of evaluation metrics crucial for objectively assessing the factual consistency and grounding capabilities of LLMs in this domain. Finally, Nayeon et al. [18] evaluate LLM factual consistency in Data-to-Text Generation (DTG), noting that while larger models generally perform better, source-reference divergence can lead to unreliable generation, impacting factual consistency.

B. Self-Correction and Advanced Prompting Techniques for LLMs

This subsection explores recent research on self-correction mechanisms and advanced prompting techniques for Large Language Models (LLMs), crucial for enhancing their reasoning capabilities and reliability. Qingjie et al. [19] investigate the limitations and potential pitfalls of intrinsic self-correction, revealing how it can lead to instability, prompt bias, and cognitive biases, particularly when oracle feedback is absent, and propose mitigation strategies. Complementing this, Zihan et al. [20] provide a comprehensive analysis of factors influencing the effectiveness of **Chain-of-Thought (CoT) prompting**, offering guidance on its application and future research directions for optimizing LLM performance. Further theoretical insights into LLM self-correction from an in-context learning perspective are offered by Yuchen et al. [21] and Yifei et al. [22], highlighting how architectural components like softmax and multi-head attention contribute to refined responses and spontaneous step-level self-correction, especially in complex reasoning tasks like mathematics. These findings align with broader efforts to understand and enhance LLM generalization and multi-capability reasoning [23]. Building on prompting advancements, Jun et al. [24] introduce "Hypothesis Testing Prompting," a novel approach to improve intermediate reasoning in LLMs by incorporating conclusion assumptions and backward reasoning, directly addressing issues of invalid reasoning paths encountered in techniques like CoT, particularly relevant for **zero-shot learning** scenarios. Moreover, Loka et al. [25] demonstrate that careful **prompt engineering**, specifically using "fair prompts" and a zero temperature setting, can unlock effective intrinsic self-correction in LLMs. Liangming et al. [26] offer a comprehensive overview of self-correction strategies, including techniques leveraging automated feedback for **critique and refinement** of model outputs. However, Jie et al. [27] highlight limitations in current LLM self-correction and advanced prompting for reasoning in multimodal contexts, demonstrating that techniques like Chain-of-Thought prompting fall short in complex multimodal tasks, indicating a critical gap in achieving robust multimodal reasoning capabilities. However, recent advancements also explore visual in-context learning for large vision-language models, extending such capabilities to multi-modal inputs [28].

III. Method

Our proposed framework, **"Context-Grounded Factuality Enhancement in LLM Responses via Multi-Stage Critique and Refinement"**, is meticulously designed to automatically improve the factual consistency and contextual grounding of Large Language Model (LLM) generated content. This is achieved by enabling LLMs to act as sophisticated self-critiquing and self-revising agents, operating in a zero-shot manner without requiring any additional model training or fine-tuning. The core idea is to leverage the inherent reasoning and instruction-following capabilities of powerful pre-trained LLMs to systematically identify and correct factual hallucinations and ungrounded information, thereby enhancing the reliability of their outputs.

A. Problem Formulation

Given an input question Q, which represents the user's query or instruction, and a set of relevant source documents $D = \{d_1, d_2, \ldots, d_n\}$, which provide the necessary factual context, an initial response R_0 is first generated by a base LLM. The objective of our framework is to produce a refined response R' that is factually consistent with the information explicitly provided in D and contextually grounded to both Q and D. This process aims to significantly minimize factual hallucinations (statements not supported by D) and contextual detachment (information irrelevant or not directly derivable from Q and D) present in the initial response R_0. This transformation process can be formally represented as a function F that maps the input tuple (Q, D, R_0) to the desired refined output R':

$$R' = F(Q, D, R_0) \tag{1}$$

Our method orchestrates this complex transformation by employing a single, powerful LLM to fulfill distinct yet complementary roles: that of a "Fact Verifier" and a "Content Reviser," operating in a structured, multi-stage process.

B. Multi-Stage Critique and Refinement Framework

The proposed framework conceptually operates in two principal stages: a critique phase dedicated to comprehensive error identification and a subsequent refinement phase focused on precise content revision. Both stages are executed by the same underlying LLM, which is dynamically guided by meticulously designed prompts to assume the appropriate role and perform the required task.

1) Initial Response Generation: The process commences with obtaining an initial response, denoted as R_0. This preliminary response is generated by a designated base LLM (e.g., models such as GPT-3.5-turbo or Llama 3 70B) given the input question Q and the set of source documents D. This initial generation serves as the primary input for our subsequent critique and refinement mechanism. The generation of R_0 can be formally expressed as:

$$R_0 = \mathcal{M}_{\text{base}}(Q, D) \tag{2}$$

where $\mathcal{M}_{\text{base}}$ represents the base Large Language Model responsible for generating the initial response.

2) Critique Phase: Fact Verification and Error Identification: In this pivotal phase, the chosen LLM assumes the role of a **"Fact Verifier"**. It rigorously analyzes the initial response R_0 by comparing its claims against the provided original question Q and the source documents D. The LLM is specifically prompted to systematically scrutinize R_0 for any factual inconsistencies, logical flaws, or statements that are not directly supported by or contradict the information within D. The output of this phase is a detailed critique, denoted as C_E, which is a structured representation of identified errors. This typically includes the specific erroneous statements, their classified types (e.g., factual hallucination, ungrounded claim, contradiction, missing information), and precise supporting evidence or counter-evidence from D for the identified discrepancies. This stage can be conceptually represented as:

$$C_E = \mathcal{M}_{\text{verifier}}(P_{\text{verify}}, Q, D, R_0) \tag{3}$$

Here, $\mathcal{M}_{\text{verifier}}$ refers to the LLM operating in its fact-checking and verification role, and P_{verify} is the specific prompt template engineered to elicit this critical analysis and structured error report.

3) Refinement Phase: Content Revision: Following the comprehensive critique, the LLM seamlessly transitions into its **"Content Reviser"** role. Leveraging the initial response R_0, the original question Q, the source documents D, and crucially, the self-generated detailed critique C_E, the LLM is prompted to produce a revised and improved response R'. The primary objective of this phase is to meticulously correct all identified errors, address ungrounded claims, and ensure the final output is factually accurate, coherent, and contextually anchored to D. The generation of the refined response can be formally expressed as:

$$R' = \mathcal{M}_{\text{reviser}}(P_{\text{revise}}, Q, D, R_0, C_E) \tag{4}$$

In this equation, $\mathcal{M}_{\text{reviser}}$ denotes the LLM operating in its revision and generation role, and P_{revise} is the prompt specifically designed to guide the content revision process. It is paramount to emphasize that $\mathcal{M}_{\text{verifier}}$ and $\mathcal{M}_{\text{reviser}}$ typically refer to the **same** underlying Large Language Model. Their distinct roles and functionalities are purely distinguished and invoked by the specific instructions and contextual cues embedded within the respective prompts, P_{verify} and P_{revise}.

C. Zero-Shot Prompting Strategies

A cornerstone of our methodological approach is its exclusive reliance on **zero-shot prompting**. This implies that no specific fine-tuning, additional training, or dataset-specific adaptation is performed on the LLMs employed within our framework. Instead, we harness the inherent, pre-trained capabilities of these powerful models to follow complex instructions, perform in-context reasoning, and generate high-quality outputs directly. We explore two distinct prompting strategies to guide the LLM through the critique and refinement process:

1) Simple Prompting: For the simple prompting strategy, a concise and direct instruction is provided to the LLM. This approach implicitly combines the critique and refinement steps into a single, straightforward command within the prompt. The LLM is directly asked to analyze the initial response R_0 against the provided context D and question Q, then immediately produce a corrected version R'. While effective for simpler cases, this strategy does not explicitly guide the LLM through intermediate reasoning steps. The generation of the refined response R' under this strategy can be formulated as:

$$R' = \mathcal{M}(\mathcal{P}_{\text{simple}}(Q, D, R_0)) \tag{5}$$

where $\mathcal{M}$ represents the general Large Language Model, and $\mathcal{P}_{\text{simple}}$ is the prompt string containing the direct, consolidated instruction for error identification and correction.

2) Chain-of-Thought (CoT) Guided Prompting: Our more advanced and robust strategy employs a multi-stage **Chain-of-Thought (CoT)** approach, significantly enhancing the LLM's ability to perform explicit reasoning and structured self-correction. This strategy is meticulously implemented by crafting a complex and detailed prompt, $\mathcal{P}_{\text{CoT}}$, that explicitly guides the LLM through a sequence of internal, verifiable steps. This structured reasoning process enables the LLM to articulate its thought process, making its critique and refinement more transparent and effective. The steps outlined within the CoT prompt typically include:

1) **Verification Steps Listing:** The LLM is first explicitly instructed to outline the systematic steps it will undertake to verify the factual consistency and contextual grounding of the initial response R_0 against the source documents D. This encourages a methodical approach.
2) **Error Identification and Classification:** Subsequently, the LLM is prompted to meticulously pinpoint specific errors or ungrounded statements within R_0. It is required to classify these errors by type (e.g., factual inaccuracy, ungrounded information, logical inconsistency) and, critically, to provide concrete, direct evidence from D to support its claims regarding the discrepancies. The output of this step constitutes the detailed critique C_E.
3) **Response Revision:** Finally, based on its self-generated critique C_E and its comprehensive understanding of Q and D, the LLM is instructed to produce the final, revised response R'. This revision must address all identified issues and ensure factual accuracy and contextual relevance.

While presented as distinct conceptual stages in Section 2.2 for clarity, for the strategy, the entire process (from critique generation to final refinement) is orchestrated by a single, comprehensive CoT prompt. The LLM's output for this strategy often explicitly includes the intermediate reasoning steps (the critique C_E) before presenting the final refined response R'. This structured reasoning process allows the LLM to more thoroughly analyze, diagnose, and correct its outputs, leading to superior performance as demonstrated in our experimental results. The output of this strategy, encompassing both the critique and the refined response, can be formally represented as:

$$(C_E, R') = \mathcal{M}(\mathcal{P}_{\mathrm{CoT}}(Q, D, R_0)) \tag{6}$$

Here, $\mathcal{P}_{\mathrm{CoT}}$ represents the elaborately designed prompt that elicits this multi-stage, explicit reasoning process from the general LLM $\mathcal{M}$.

D. Data Preparation and Processing

For each instance utilized in our evaluation, specifically from datasets such as HotpotQA and ELI5, the input data is meticulously standardized into a tuple (Q, D, R_0). Q consistently represents the input question or prompt provided to the LLM. D comprises the set of relevant source documents or context passages, which are indispensable for factual verification and grounding of the LLM's response. R_0 denotes the initial, uncorrected response generated by a base LLM prior to our framework's intervention. We ensure that the provided set of documents D contains sufficient and verifiable information to rigorously assess the factual claims made in R_0 and to identify any statements that are not grounded within the provided context. Crucially, no specific pre-processing of the text content beyond basic tokenization and formatting (e.g., concatenation of documents into a single string if necessary) for optimal LLM input is required. This aligns perfectly with our zero-shot methodology, where the efficacy of the framework hinges entirely on the sophistication of prompt engineering to effectively guide the LLM, rather than on data-specific feature engineering or extensive pre-computation.

IV. Experiments

In this section, we detail the experimental setup, evaluate the performance of our proposed framework, and compare it against various baselines. We further analyze the impact of different prompting strategies and present results from human evaluations to corroborate our findings.

A. Experimental Setup

1) Models: Our framework is designed to be compatible with various large language models (LLMs) serving as the core "Fact Verifier" and "Content Reviser." For our experiments, we selected a diverse set of representative models to assess their performance within our framework:

- **GPT-3.5-turbo**: A widely used and capable model from OpenAI.
- **Gemini Pro**: Google's advanced multimodal model, evaluated here for its text generation capabilities.
- **Mixtral 8x7B**: A high-quality sparse mixture-of-experts model, known for its efficiency and strong performance.
- **Llama 3 70B**: A powerful open-source model, serving as the foundation for our primary proposed method.

For each model, we evaluated two distinct prompting strategies:

- **(Simple Prompting)**: This refers to a more concise and direct prompt, implicitly requesting the model to perform verification and correction without explicit intermediate reasoning steps, as described in Section 2.4.1.
- **(Chain-of-Thought Guided Prompting)**: This signifies a more complex and elaborate prompt designed to elicit multi-stage Chain-of-Thought (CoT) reasoning, guiding the model through explicit verification steps, detailed error identification, and structured revision, as detailed in Section 2.4.2.

Our proposed method, **Ours**, specifically utilizes the Llama 3 70B model with both and prompting strategies.

2) Datasets: To comprehensively evaluate the models' abilities in enhancing factual consistency and contextual grounding, we employed two challenging and widely recognized datasets:

- **HotpotQA**: A multi-hop question answering dataset that requires models to aggregate information from multiple supporting documents to formulate an answer. This dataset is particularly effective for evaluating a model's capacity to process complex, distributed information and anchor its responses firmly within the provided context, thereby minimizing ungrounded statements.
- **ELI5 (Explain Like I'm 5)**: A question answering dataset focused on generating simplified explanations of complex concepts. We specifically utilized its longer answer segments to assess the models' proficiency in maintaining explanatory accuracy and coherence without introducing spurious information or factual inaccuracies.

For both datasets, input data was standardized to '(Question, Relevant Source Documents, Initial LLM Response)', where the initial response was generated by a base LLM prior to our framework's application.

3) Evaluation Metrics: We assessed the performance using two primary quantitative metrics, both ranging from 0 to 1, with higher scores indicating better performance:

- **Fact Consistency Score (FCS)**: Measures the factual accuracy of the generated response relative to the provided source documents. It quantifies the degree to which statements in the response are supported by or consistent with the ground truth information.
- **Context Grounding Score (CGS)**: Evaluates how well the generated response is anchored to the provided context and question. It penalizes information that is irrelevant, ungrounded, or deviates from the scope defined by the input.

These metrics were computed automatically using established evaluation protocols for factual accuracy and groundedness in LLM outputs.

B. Baselines

We compared our framework against several strong baselines to demonstrate its effectiveness:

- **Base LLM (No Correction)**: Represents the raw output of a large language model (e.g., Llama 3 70B) without any post-generation critique or refinement. This baseline highlights the inherent limitations of LLMs in terms of factual consistency and contextual grounding.
- **Prior Work (Simple Self-Correction)**: Encompasses existing simple self-correction strategies that involve a single-pass or less structured prompting approach for minor revisions. This baseline reflects the current state-of-the-art in straightforward self-correction mechanisms.
- **Other LLMs with and strategies**: We also present the performance of GPT-3.5-turbo, Gemini Pro, and Mixtral 8x7B, each employing both simple and Chain-of-Thought prompting strategies. These serve as strong comparative models, illustrating the general applicability and benefits of our multi-stage approach across different LLM architectures.

C. Main Results

Table 1 summarizes the performance of our proposed method and all baseline models on the HotpotQA and ELI5 datasets across FCS and CGS metrics.

Table 1. Model Comparison Results on HotpotQA and ELI5 Datasets.

Model/Method	**HotpotQA (FCS)**	**HotpotQA (CGS)**	**ELI5 (FCS)**	**ELI5 (CGS)**
Base LLM (No Correction)	0.62	0.58	0.55	0.50
Prior Work (Simple Self-Correction)	0.68	0.65	0.60	0.57
GPT-3.5-turbo_s	0.69	0.67	0.61	0.59
GPT-3.5-turbo_b	0.72	0.70	0.65	0.62
Gemini Pro_s	0.71	0.69	0.64	0.61
Gemini Pro_b	0.75	0.73	0.68	0.65
Mixtral 8x7B_s	0.74	0.72	0.67	0.64
Mixtral 8x7B_b	0.77	0.75	0.70	0.68
Ours (Llama 3 70B_s)	**0.78**	**0.76**	**0.71**	**0.69**
Ours (Llama 3 70B_b)	**0.81**	**0.80**	**0.75**	**0.73**

a) Analysis of Results: The results in Table 1 clearly demonstrate the effectiveness of our proposed framework.

- **Baseline Performance:** The "Base LLM (No Correction)" shows the inherent limitations of raw LLM outputs, with relatively low FCS and CGS scores, highlighting the prevalence of factual hallucinations and ungrounded information.
- **Existing Solutions:** "Prior Work (Simple Self-Correction)" offers a modest improvement over the base LLM, indicating that simple correction mechanisms can provide some benefit, but significant room for improvement remains.
- **CoT's Advantage:** A crucial observation is the consistent performance boost achieved by employing the Chain-of-Thought (CoT) guided prompting strategy compared to the simple prompting strategy for all evaluated models. For instance, GPT-3.5-turbo_b consistently outperforms GPT-3.5-turbo_s across all metrics and datasets. This empirically validates our hypothesis that enabling LLMs to perform explicit reasoning and structured verification steps significantly enhances their ability to identify and correct errors.
- **Our Method's Superiority:** Our proposed "Ours" method, leveraging the powerful Llama 3 70B model, consistently achieves the best performance across all metrics and datasets. Particularly, "Ours (Llama 3 70B_b)" sets new state-of-the-art results, with FCS scores of 0.81 on HotpotQA and 0.75 on ELI5, and similarly high CGS scores. This superior performance underscores the efficacy of our multi-stage critique and refinement framework when combined with a strong underlying LLM and the strategic application of CoT prompting. It demonstrates that guiding the LLM through a structured verification and revision process, without requiring additional training, is highly effective in mitigating factual hallucinations and improving contextual grounding.

D. Analysis of Prompting Strategies

To further elucidate the impact of our zero-shot prompting strategies, we conducted an in-depth analysis of the performance differences between the simple and Chain-of-Thought approaches. As observed in Table 1, the strategy consistently yields higher FCS and CGS scores than the strategy for every LLM tested (GPT-3.5-turbo, Gemini Pro, Mixtral 8x7B, and Llama 3 70B).

This consistent improvement can be attributed to several factors inherent in the CoT approach:

- **Explicit Reasoning Path:** The CoT prompts compel the LLM to articulate its reasoning process step-by-step. This explicit decomposition of the task (e.g., identify verification steps, pinpoint errors, classify error types, provide evidence) allows the model to engage in deeper, more structured analysis of the initial response against the provided context.
- **Enhanced Error Identification:** By requiring the LLM to list specific errors and provide supporting evidence from the source documents, the strategy forces a more thorough and precise identification of factual inconsistencies and ungrounded statements. This reduces the likelihood of overlooking subtle errors.

- **Targeted Correction:** With a clear and detailed critique (C_E) generated in the preceding step, the LLM in its "Content Reviser" role can perform more targeted and accurate corrections. The explicit error types and evidence guide the revision process more effectively than a general instruction.
- **Reduced Ambiguity:** The structured nature of CoT prompts reduces ambiguity in the task instructions, leading to more reliable and consistent performance, especially for complex cases where factual nuances or multiple pieces of context need to be reconciled.

The results strongly suggest that investing in meticulously designed CoT prompts is a highly effective, zero-shot approach to leveraging the intrinsic reasoning capabilities of LLMs for self-correction and factuality enhancement.

E. Human Evaluation

While automatic metrics provide quantitative insights, human evaluation offers a crucial qualitative assessment of response quality, particularly concerning nuances of factual consistency, contextual relevance, and overall coherence that automated scores might miss. We conducted a human evaluation on a random subset of 100 responses from each dataset (HotpotQA and ELI5) for three key methods: the Base LLM (No Correction), Prior Work (Simple Self-Correction), and our best performing method, Ours (Llama 3 70B_b).

Three independent expert annotators, blinded to the model origin, rated each response on a 5-point Likert scale (1=Poor, 5=Excellent) for the following criteria:

- **Factual Accuracy**: How well the response aligns with the facts presented in the source documents.
- **Contextual Grounding**: How well the response uses and stays relevant to the provided context.
- **Fluency and Coherence**: The readability, grammatical correctness, and logical flow of the response.
- **Overall Quality**: A holistic judgment of the response's utility and correctness.

The average scores are presented in Table 2. Inter-annotator agreement was calculated using Fleiss' Kappa, yielding a moderate agreement of $\kappa = 0.68$.

Table 2. Human Evaluation Results (Average Likert Score, 1-5).

Model/Method	Factual Accuracy	Contextual Grounding	Fluency & Coherence	Overall Quality
Base LLM (No Correction)	2.8	2.9	4.2	2.7
Prior Work (Simple Self-Correction)	3.5	3.4	4.3	3.3
Ours (Llama 3 70B_b)	**4.6**	**4.5**	**4.5**	**4.4**

a) Analysis of Human Evaluation: The human evaluation results strongly corroborate the findings from our automatic metrics.

- Our method (Ours, Llama 3 70B_b) received significantly higher scores across all evaluation criteria, particularly in Factual Accuracy and Contextual Grounding, confirming its superior ability to mitigate hallucinations and ungrounded statements.
- While Base LLM outputs were often fluent, their low scores in factual accuracy and grounding highlight the critical need for correction mechanisms.
- Prior work showed improvement, but our multi-stage critique and refinement framework was perceived by human annotators as producing responses that are not only factually correct and well-grounded but also maintain high fluency and overall quality, making them highly reliable and useful.

These human judgments provide strong evidence that our framework successfully addresses the core challenges of factual consistency and contextual grounding in LLM-generated content.

F. Ablation Studies

To systematically understand the contribution of each key component within our "Context-Grounded Factuality Enhancement" framework, we conducted ablation studies using the Llama 3 70B model on both datasets. We specifically investigate the impact of the explicit critique generation (C_E) and the structured multi-stage roles.

a) Analysis of Ablation Results: Table 3 presents the performance of our full framework compared to its ablated variants.

Table 3. Ablation Study Results on HotpotQA and ELI5 Datasets.

Method	HotpotQA (FCS)	HotpotQA (CGS)	ELI5 (FCS)	ELI5 (CGS)
Ours (Llama 3 70B_b) - Full Framework	**0.81**	**0.80**	**0.75**	**0.73**
Ours (No Explicit Critique Report)	0.79	0.77	0.72	0.70
Ours (Single-Pass CoT, No Distinct Roles)	0.76	0.74	0.69	0.67

- **Impact of Explicit Critique Report (C_E):** "Ours (No Explicit Critique Report)" refers to a setup where the LLM is guided by a CoT prompt to perform verification and revision steps internally, but it does not explicitly output the detailed critique C_E as a separate, structured report before generating R'. While still performing well due to the internal CoT reasoning, its performance is slightly lower than the full framework (e.g., 0.79 FCS on HotpotQA vs. 0.81). This indicates that the act of **explicitly generating** the critique C_E forces the LLM to formalize its error identification and evidence gathering, leading to a more robust and accurate subsequent revision. The structured C_E acts as a concrete intermediate representation that the LLM can more reliably reference for refinement.
- **Impact of Distinct Roles and Multi-Stage Process:** "Ours (Single-Pass CoT, No Distinct Roles)" represents a scenario where a single, complex CoT prompt is used to guide the LLM to perform verification and revision, but without explicitly instructing it to assume the distinct "Fact Verifier" and "Content Reviser" roles, and without a clear conceptual separation into critique and refinement stages as outlined in Section 2.2. The performance drop (e.g., 0.76 FCS on HotpotQA vs. 0.81 for the full framework) highlights the importance of our meticulously designed multi-stage framework. By defining clear roles and sequential steps, the framework effectively decomposes a complex problem into manageable sub-problems, guiding the LLM to focus on specific tasks (verification, then revision) in a structured manner. This structured guidance minimizes the cognitive load on the LLM, enabling it to execute each phase more effectively.

These ablation studies empirically confirm that both the explicit generation of a detailed critique report and the conceptual separation into distinct "Fact Verifier" and "Content Reviser" roles operating in a multi-stage process are crucial for the superior performance of our proposed framework. Each component contributes significantly to enhancing the LLM's ability to self-critique and refine its outputs for factual consistency and contextual grounding.

G. Qualitative Analysis and Error Correction Examples

To provide deeper insights into how our framework operates, particularly the 'Ours (Llama 3 70B_b)' method, we present a qualitative analysis of typical errors identified and corrected. Our framework excels at addressing two primary categories of errors: factual hallucinations and ungrounded statements. Factual hallucinations involve statements that directly contradict or are not supported by the provided source documents. Ungrounded statements include information that is plausible but not derivable from the given context, or content that deviates from the question's scope.

Table 4 illustrates a representative example from the HotpotQA dataset, showcasing the initial problematic response from a Base LLM, the detailed critique generated by our framework's "Fact Verifier" role, and the final refined response produced by the "Content Reviser."

Table 4. Case Study: Error Correction Example (HotpotQA).

Component	Content
Question (Q)	Who wrote the novel "The Hitchhiker's Guide to the Galaxy" and what year was the first book published?
Source Documents (D)	**Document 1:** "The Hitchhiker's Guide to the Galaxy is a comedy science fiction series created by Douglas Adams. Originally a 1978 radio comedy broadcast on BBC Radio 4, it was later adapted to other formats, including a 'trilogy of five' books that sold over 15 million copies. The first book, also titled 'The Hitchhiker's Guide to the Galaxy', was published in 1979." **Document 2:** "Douglas Adams (1952-2001) was an English author, satirist, and dramatist. He is best known as the author of The Hitchhiker's Guide to the Galaxy, which originated in 1978 as a BBC radio series and was adapted into a book series."
Initial Response (R_0) from Base LLM	The novel "The Hitchhiker's Guide to the Galaxy" was written by Douglas Adams, and the first book was published in 1978.
Critique (C_E) from Ours	**Error 1: Factual Hallucination (Year of Publication)** **Statement:** "the first book was published in 1978." **Evidence from D:** Document 1 states: "The first book... was published in 1979." **Reasoning:** The initial response incorrectly states the publication year of the first book. The source clearly indicates 1979, not 1978.
Refined Response (R') from Ours	The novel "The Hitchhiker's Guide to the Galaxy" was written by Douglas Adams, and the first book was published in 1979.

As demonstrated in Table 4, the Base LLM, despite being provided with the correct documents, generated a factual hallucination regarding the publication year. Our framework, in its "Fact Verifier" role, meticulously identified this specific error, classified it as a "Factual Hallucination," and provided precise counter-evidence from Document 1. This detailed critique (C_E) then served as a clear directive for the "Content Reviser" role, enabling it to produce a factually accurate and fully grounded refined response (R'). This level of explicit error identification and evidence-based correction is a key differentiator of our approach, leading to the significant improvements observed in both automatic and human evaluations.

H. Computational Cost Analysis

While our multi-stage critique and refinement framework significantly enhances factual consistency and contextual grounding, it inherently involves more computational steps compared to a single-pass generation. This translates to increased token usage and, consequently, longer inference times.

Table 5 provides an overview of the average token counts for input and output across different strategies for the Llama 3 70B model. The total tokens processed directly correlate with the computational cost (e.g., API calls, GPU time).

Table 5. Average Token Usage per Query for Llama 3 70B.

Method	Avg. Input Tokens	Avg. Output Tokens	Avg. Total Tokens
Base LLM (No Correction)	500	150	650
Ours (Llama 3 70B_s)	600	180	780
Ours (Llama 3 70B_b)	750	300	1050

a) Analysis of Computational Cost:

doi:10.20944/preprints202508.0700.v1

- **Increased Input Context:** Both and especially strategies require a larger input context compared to the Base LLM. This is because the prompts for self-correction include the original question, source documents, and the initial response, along with detailed instructions. The strategy, with its comprehensive CoT prompt, naturally has a larger average input token count (750 tokens) than (600 tokens) and the Base LLM (500 tokens).
- **Higher Output Tokens for CoT:** The most significant increase in token usage comes from the output of the Chain-of-Thought strategy. Since it explicitly generates the detailed critique (C_E) before the refined response (R'), the average output token count is considerably higher (300 tokens) compared to (180 tokens) and the Base LLM (150 tokens). This additional output is the explicit reasoning process that underpins the improved performance.
- **Trade-off between Performance and Cost:** The "Ours (Llama 3 70B_b)" method, while achieving the highest performance, also incurs the highest computational cost, approximately 1.6 times that of the Base LLM in terms of total tokens. This represents a clear trade-off: the enhanced factual consistency and contextual grounding come at the expense of increased inference time and API costs (for commercial models).
- **Practical Implications:** For applications where extreme low-latency is critical, a simpler approach or even accepting the limitations of a Base LLM might be considered. However, for use cases demanding high factual accuracy and reliability, such as knowledge base construction, automated content generation in sensitive domains, or critical decision support systems, the increased computational cost of our framework is a justifiable investment given the substantial gains in output quality. Future work could explore optimizations to reduce token usage while retaining the benefits of CoT reasoning, perhaps through more concise critique formats or distillation techniques.

V. Conclusion

In this paper, we have presented "Context-Grounded Factuality Enhancement in LLM Responses via Multi-Stage Critique and Refinement," a novel and effective automated framework designed to mitigate the pervasive issues of factual hallucinations and contextual detachment in Large Language Model (LLM) generated content. Recognizing the critical need for reliable LLM outputs, especially in high-stakes domains, our approach empowers LLMs to act as sophisticated self-critiquing and self-revising agents.

Our core contribution lies in the development of a zero-shot prompting methodology that orchestrates a multi-stage process, conceptualizing the LLM's role as a distinct "Fact Verifier" and "Content Reviser." By meticulously guiding the LLM through Chain-of-Thought (CoT) reasoning, our framework enables explicit error identification, classification, and evidence-based correction. This structured approach, which avoids any additional training or fine-tuning of the underlying LLMs, leverages their innate instruction-following and reasoning capabilities to deliver enhanced factual consistency and contextual grounding.

Our comprehensive experimental evaluation on the HotpotQA and ELI5 datasets unequivocally demonstrated the superior performance of our proposed method. We observed significant improvements in both Fact Consistency Score (FCS) and Context Grounding Score (CGS) compared to base LLM outputs and existing simple self-correction techniques. The consistent and substantial performance gains achieved by our CoT-guided prompting strategy (denoted) across various LLMs, including GPT-3.5-turbo, Gemini Pro, Mixtral 8x7B, and particularly Llama 3 70B, empirically validated the efficacy of explicit, multi-step reasoning in error correction. Furthermore, human evaluations provided crucial qualitative evidence, confirming that our framework produces responses that are not only factually accurate and well-grounded but also maintain high levels of fluency and overall quality. Ablation studies further solidified our findings, proving the indispensable contributions of both explicit critique generation and the conceptual separation of distinct roles within the multi-stage framework.

While our method achieves state-of-the-art results, we acknowledge the inherent trade-off with computational cost. The multi-stage reasoning and explicit critique generation lead to increased token usage and, consequently, longer inference times compared to single-pass generation. This highlights a critical area for future research.

For future work, we plan to explore several avenues. Firstly, we aim to investigate optimizations for reducing computational overhead, potentially through more concise critique formats, knowledge distillation techniques, or adaptive prompting strategies that dynamically adjust the level of CoT reasoning based on task complexity or confidence scores. Secondly, extending this framework to more complex reasoning tasks, such as logical inference or summarization with strict factual constraints, and to other modalities like multimodal generation, presents exciting opportunities. Thirdly, we intend to evaluate the framework's robustness across a wider array of diverse domains and languages to ensure its generalizability. Ultimately, our work represents a significant step towards building more reliable and trustworthy LLM-powered applications, paving the way for their responsible deployment in sensitive and high-stakes environments where factual accuracy is paramount.

References

1. Zhou, Y.; Tao, W.; Zhang, W. Triple sequence generative adversarial nets for unsupervised image captioning. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7598–7602.
2. Wang, Q.; Hu, H.; Zhou, Y. Memorymamba: Memory-augmented state space model for defect recognition. *arXiv preprint arXiv:2405.03673* **2024**.
3. Cao, M.; Dong, Y.; Cheung, J.C.K. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022, pp. 3340–3354. https://doi.org/10.18653/V1/2022.ACL-LONG.236.
4. Zhou, Y.; Song, L.; Shen, J. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. *arXiv preprint arXiv:2506.19835* **2025**.
5. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.
6. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics, 2018, pp. 2369–2380. https://doi.org/10.18653/V1/D18-1259.
7. Fan, A.; Jernite, Y.; Perez, E.; Grangier, D.; Weston, J.; Auli, M. ELI5: Long Form Question Answering. In Proceedings of the Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019, pp. 3558–3567. https://doi.org/10.18653/V1/P19-1346.
8. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *CoRR* **2023**. https://doi.org/10.48550/ARXIV.2311.05232.
9. Talukdar, W.; Biswas, A. Improving Large Language Model (LLM) fidelity through context-aware grounding: A systematic approach to reliability and veracity. *CoRR* **2024**. https://doi.org/10.48550/ARXIV.2408.04023.
10. Zhou, Y.; Shen, T.; Geng, X.; Long, G.; Jiang, D. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2559–2575.

11. Zhou, Y.; Geng, X.; Shen, T.; Long, G.; Jiang, D. Eventbert: A pre-trained model for event correlation reasoning. In Proceedings of the Proceedings of the ACM Web Conference 2022, 2022, pp. 850–859.
12. Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G.L.; Corney, D.P.A.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.Y.; et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mac. Intell.* **2024**, pp. 852–863. https://doi.org/10.1038/S42256-024-00881-Z.
13. Huang, Y.; Huang, J. A Survey on Retrieval-Augmented Text Generation for Large Language Models. *CoRR* **2024**. https://doi.org/10.48550/ARXIV.2404.10981.
14. Yang, L.; Chen, H.; Li, Z.; Ding, X.; Wu, X. Give us the Facts: Enhancing Large Language Models With Knowledge Graphs for Fact-Aware Language Modeling. *IEEE Trans. Knowl. Data Eng.* **2024**, pp. 3091–3110. https://doi.org/10.1109/TKDE.2024.3360454.
15. Zhou, Y.; Geng, X.; Shen, T.; Pei, J.; Zhang, W.; Jiang, D. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* **2021**.
16. Ghosh, B.; Hasan, S.; Arafat, N.A.; Khan, A. Logical Consistency of Large Language Models in Fact-Checking. In Proceedings of the The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net, 2025.
17. Kamalloo, E.; Dziri, N.; Clarke, C.L.A.; Rafiei, D. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. Association for Computational Linguistics, 2023, pp. 5591–5606. https://doi.org/10.18653/V1/2023.ACL-LONG.307.
18. Lee, N.; Ping, W.; Xu, P.; Patwary, M.; Fung, P.; Shoeybi, M.; Catanzaro, B. Factuality Enhanced Language Models for Open-Ended Text Generation. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.
19. Zhang, Q.; Wang, D.; Qian, H.; Li, Y.; Zhang, T.; Huang, M.; Xu, K.; Li, H.; Yan, L.; Qiu, H. Understanding the Dark Side of LLMs' Intrinsic Self-Correction. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025. Association for Computational Linguistics, 2025, pp. 27066–27101.
20. Yu, Z.; He, L.; Wu, Z.; Dai, X.; Chen, J. Towards Better Chain-of-Thought Prompting Strategies: A Survey. *CoRR* **2023**. https://doi.org/10.48550/ARXIV.2310.04959.
21. Yan, Y.; Jiang, J.; Liu, Y.; Cao, Y.; Xu, X.; Zhang, M.; Cai, X.; Shao, J. S^3cMath: Spontaneous Step-Level Self-Correction Makes Large Language Models Better Mathematical Reasoners. In Proceedings of the AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. AAAI Press, 2025, pp. 25588–25596. https://doi.org/10.1609/AAAI.V39I24.34749.
22. Wang, Y.; Wu, Y.; Wei, Z.; Jegelka, S.; Wang, Y. A Theoretical Understanding of Self-Correction through In-context Alignment. In Proceedings of the Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024.
23. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
24. Sun, J.; Pan, Y.; Yan, X. Improving intermediate reasoning in zero-shot chain-of-thought for large language models with filter supervisor-self correction. *Neurocomputing* **2025**, p. 129219. https://doi.org/10.1016/J.NEUCOM.2024.129219.
25. Li, L.; Chen, G.; Su, Y.; Chen, Z.; Zhang, Y.; Xing, E.P.; Zhang, K. Confidence Matters: Revisiting Intrinsic Self-Correction Capabilities of Large Language Models. *CoRR* **2024**. https://doi.org/10.48550/ARXIV.2402.12563.
26. Pan, L.; Saxon, M.; Xu, W.; Nathani, D.; Wang, X.; Wang, W.Y. Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies. *CoRR* **2023**. https://doi.org/10.48550/ARXIV.2308.03188.

27. Huang, J.; Chen, X.; Mishra, S.; Zheng, H.S.; Yu, A.W.; Song, X.; Zhou, D. Large Language Models Cannot Self-Correct Reasoning Yet. In Proceedings of the The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024.
28. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.