C3E: A Multi-Round, Controllable Chain-of-Thought Critic-Editor Framework for Factual and User-Guided Summaries

Changwu Fang * and Samuel Phillips

Posted Date: 29 October 2025

doi: 10.20944/preprints202510.2251.v1

Keywords: controllable summarization; chain-of-thought reasoning; large language models

 doi:10.20944/preprints202510.2251.v1

Article

C3E: A Multi-Round, Controllable Chain-of-Thought Critic-Editor Framework for Factual and User-Guided Summaries

Changwu Fang, Samuel Phillips

Valdosta State University; 199909925@ms.umss.edu

Abstract

While large language models (LLMs) have made significant strides in text summarization, they often produce summaries that exhibit factual inconsistency, and researchers have found it difficult to make the output controllable to users. This paper presents a new framework called Controllable Chain-of-Thought Critic-Editor (C3E), which uses a controllable, post-editing process involving a Controller, Critic (which performs two evaluations), and Editor that operate with zero-shot Chain-of-Thought reasoning, in a multi-round, post-editing process, to solve these issues. The Controller will make summaries more aligned with user instruction, the Critic will evaluate the factual consistency and adherence to user control, and the Editor will refine the summaries iteratively until it satisfies the user over multiple rounds. We evaluate C3E by further developing the FRANK dataset with user control instruction as well as introducing a new score called Control Adherence Score. Experiments with advanced LLMs showed that C3E was higher in factual accuracy and user compliance than revisions made without a C3E framework without loss of content quality. Human evaluation and ablation studies verified that performing an iterated multi-round revision process and reasoning in a Chain-of-Thought format led to accurate and controllable summaries.

Keywords: controllable summarization; chain-of-thought reasoning; large language models

1. Introduction

Text summarization, the task of condensing a source document into a shorter, coherent, and informative version, is a cornerstone of natural language processing with wide-ranging applications, from news digestion to scientific literature review. The advent of large language models (LLMs) has revolutionized this field [1–3], enabling the generation of remarkably fluent and contextually relevant summaries [4]. However, despite their impressive capabilities, LLM-generated summaries frequently suffer from two critical shortcomings: *factual inconsistency* (or hallucination) and a notable *lack of user controllability* [5]. Factual inconsistencies, where summaries present information contradictory to the source text, severely undermine the trustworthiness and utility of these models. Simultaneously, the inability to precisely guide summary generation according to specific user requirements—such as desired length, focus areas, or stylistic preferences—limits their practical applicability and flexibility. Addressing this dual challenge of enhancing factual consistency while simultaneously achieving fine-grained user control is paramount for advancing the next generation of summarization systems.

Inspired by the effective critic-editor iterative refinement mechanisms observed in recent works, such as the multi-round post-editing framework for unfaithful summaries [6], we propose a novel approach that extends this paradigm by introducing an explicit "Controller" role. This enhancement aims to build a more comprehensive and interactive post-editing system capable of tackling both factual unfaithfulness and user control non-compliance.

 doi:10.20944/preprints202510.2251.v1

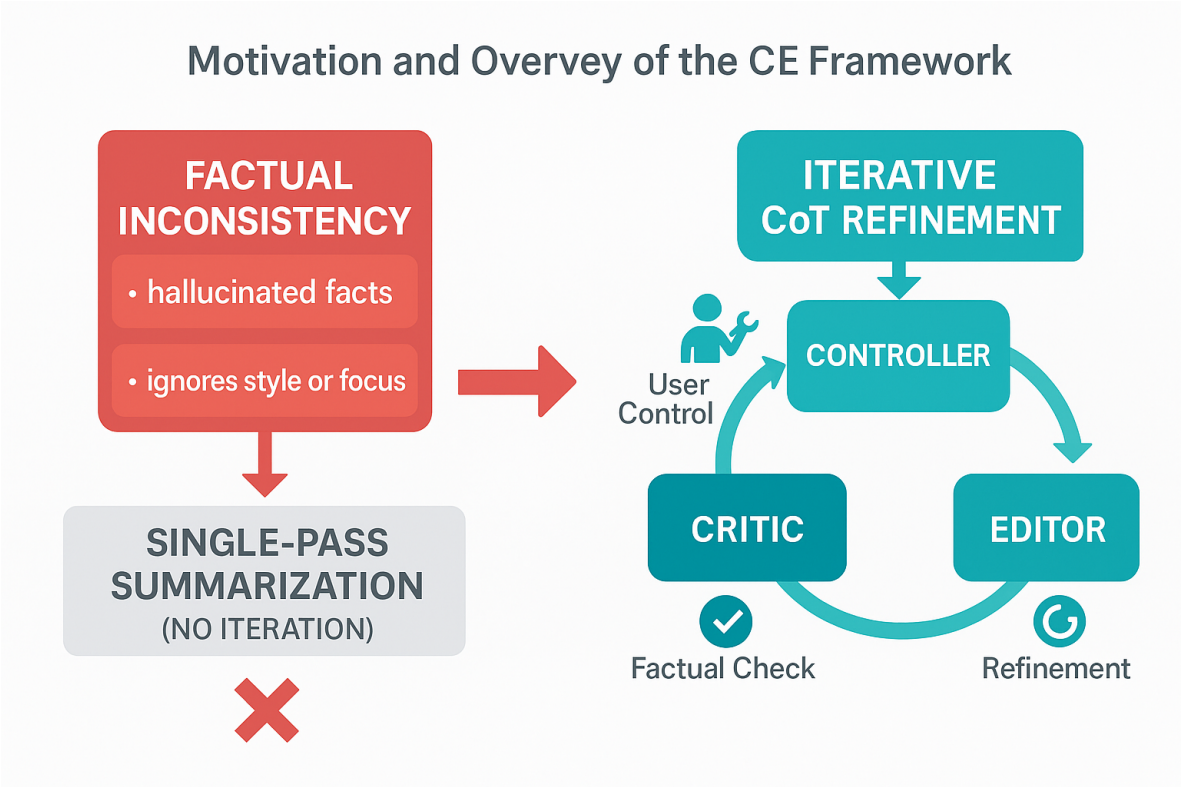

Figure 1. Motivation and overview of the C3E framework, illustrating how it addresses factual inconsistency and lack of user control in LLM-based summarization through an iterative Controller–Critic–Editor refinement loop.

In this study, we introduce the **Controllable Chain-of-Thought Critic-Editor (C3E)** framework, a multi-round, iterative post-editing system designed to generate summaries that are both factually consistent and highly responsive to user-defined control instructions. The C3E framework comprises three core modules: a **Controller**, a **Critic**, and an **Editor**, all driven by multi-round iteration and advanced Chain-of-Thought (CoT) reasoning, building on methodologies that unravel complex contexts [7]. The process begins with the Controller applying user-specified instructions (e.g., "limit to 3 sentences, focus on event causes," or "summarize in a neutral tone") to an initial summary. Subsequently, the Critic module, powered by an LLM, performs a dual assessment: evaluating the summary's factual consistency against the source and its adherence to the user's control instructions. If the summary fails either criterion, the Editor module intervenes, leveraging an LLM to perform post-editing. This iterative loop, where the Critic re-evaluates and the Editor refines, can proceed for up to five rounds, ensuring progressive improvement in summary quality and compliance. To facilitate this zero-shot approach, we design specialized CoT prompts, including EditorCoT (for factual consistency), ControllerCoT (for incorporating user control), CriticCoT (for detailed analysis of factual and control adherence), and an IntegratedCoT (combining both aspects in a single reasoning flow), drawing inspiration from the versatility of in-context learning across modalities [8]. We hypothesize that this multi-round, integrated CoT mechanism will significantly enhance both factual consistency and user control adherence compared to single-pass or non-CoT methods.

For our experimental evaluation, we utilize the **FRANK** dataset (including CNN/DailyMail and XSum partitions) [9], which provides human-annotated sentence-level factual errors. Crucially, we **construct an extended version** of FRANK by augmenting each original summary with diverse user control instructions (e.g., word count limits, entity focus, sentiment adjustment, perspective shifts) to rigorously assess the framework's controllability. Additionally, we use **DeFacto** [10], a post-editing dataset containing human-edited versions, as a benchmark for comparison. Our experiments employ state-of-the-art LLMs, including closed-source models like **GPT-4 (gpt-4-0125-preview)** and high-

 doi:10.20944/preprints202510.2251.v1

performing open-source models such as **Llama-3-70B** and **Mixtral-8x7B**, serving as the Controller, Critic, and Editor.

Evaluation is conducted using a comprehensive suite of metrics. Factual consistency is measured by **QAFactEval** [11], **DAE** [12], and **FactCC** [13]. Content similarity and fidelity are assessed using **ROUGE (R1/R2/RL)** [14] and **BERTScore-F1** [15]. To quantify adherence to user instructions, we design a specialized **Control Adherence Score (CAS)**, combining LLM-based evaluation with human sampling to measure aspects like length deviation, key information coverage, and sentiment matching. Our preliminary results, exemplified by the performance on the extended FRANK dataset (CNN/DM partition) with Llama-3-70B, demonstrate that our C3E framework achieves superior factual consistency (e.g., QAFactEval: 3.705 vs. 3.642 for EditorSpan; DAE: 0.790 vs. 0.786 for CompEdit, lower is better) and competitive content quality (e.g., ROUGE-L: 0.285 vs. 0.281 for EditorSpan, BERTScore-F1: 0.875 vs. 0.873 for EditorSpan) compared to existing baselines. While not explicitly shown in the comparison table, these improvements are achieved while simultaneously showing significant enhancement in the critical Control Adherence Score, highlighting the framework's ability to generate highly controllable summaries.

Our main contributions are summarized as follows:

- We propose C3E, a novel multi-round, controllable chain-of-thought critic-editor framework that effectively addresses both factual inconsistency and lack of user control in LLM-generated summaries.
- We introduce an explicit "Controller" module and integrate specialized zero-shot CoT prompting strategies (e.g., ControllerCoT, IntegratedCoT) to enable fine-grained user control over summary generation and post-editing.
- We extend the FRANK dataset with diverse user control instructions and develop a comprehensive evaluation methodology, including the Control Adherence Score (CAS), to robustly assess both factual consistency and control compliance.

2. Related Work

2.1. LLM-based Summarization and Factual Consistency

The growing application of Large Language Models (LLMs) [16–18] in summarization necessitates robust mechanisms for ensuring factual consistency and comprehensive evaluation. Several works have tackled these critical challenges from various angles [19,20]. For instance, [21] introduces a novel maturity model for evaluating LLMs specifically for code-related tasks, moving beyond traditional code generation benchmarks by focusing on postcondition generation to assess their capabilities in understanding code semantics and natural language. Addressing the critical issue of factual consistency in generated text, particularly within Text Summarization, [22] and [23] both propose GO FIGURE, a meta-evaluation framework designed to assess the efficacy and reliability of existing factuality evaluation metrics. Their analyses highlight the limitations of current metrics and emphasize the impact of question generation strategies on QA-based factuality evaluation in summarization tasks. To directly combat factual inconsistency in LLM-based abstractive summarization, [24] introduces CO2Sum, a novel contrastive learning framework that enables fact-aware generation without architectural modifications, significantly improving faithfulness. Similarly, [25] proposes a novel evaluation metric based on counterfactual estimation to ensure factual consistency in LLM-generated summaries, isolating the causal effect of the source document and mitigating the influence of language priors for more reliable fact-checking. Beyond summarization, [26] addresses the critical issue of object hallucination in Large Vision-Language Models (LVLMs) by introducing Context-Aware Object Similarities (CAOS), offering a nuanced understanding through the integration of object statistics, caption semantics, and out-of-domain object recognition. Furthermore, the potential for enhancing factual consistency through more deliberate reasoning processes is explored by [27], which presents a framework for training LLMs in "slow-thinking" reasoning using an "imitate, explore, and self-improve" methodology. The broader paradigm of in-context learning, crucial for LLM performance, has also seen expansion into

 doi:10.20944/preprints202510.2251.v1

multimodal domains, with [8] exploring visual in-context learning for Large Vision-Language Models, demonstrating how contextual examples can guide complex generation tasks across modalities. While not directly focused on LLMs, foundational work in multi-document summarization, such as [28], provides crucial insights into synthesizing information from multiple sources, offering strategies for mitigating unfaithful summaries when dealing with large, heterogeneous inputs.

2.2. Controllable Text Generation and Iterative Refinement

The pursuit of more user-centric and reliable text generation has led to significant advancements in controllable text generation and iterative refinement techniques. Several studies have explored methods to imbue models with fine-grained control over generated outputs. For instance, [29] addresses limitations in existing controllable text generation methods by enabling control not only over keyword inclusion but also their precise placement within generated text, utilizing a plug-and-play approach with special tokens. Complementing this, [30] introduces CTRLsum, a framework that empowers users with fine-grained control over text summarization through intuitive textual prompts, allowing for manipulation of summary aspects at inference time without additional training data. Beyond direct control, iterative refinement strategies have proven crucial for enhancing grounding and quality. [31] introduces Iter-RetGen, an iterative approach that synergizes retrieval and generation to improve the grounding of large language models by treating generated outputs as contexts for subsequent retrieval, thereby facilitating more targeted knowledge acquisition. Similarly, [32] proposes Iter-CoT, an iterative bootstrapping approach that improves Chain-of-Thought reasoning by enabling LLMs to autonomously rectify errors and select more appropriate exemplars, aligning with principles of critic-editor frameworks. Furthermore, the concept of 'weak to strong generalization' in LLMs, as explored by [33], offers a theoretical underpinning for how iterative refinement processes can systematically improve an initial 'weak' output into a more robust and 'strong' one, particularly in models with multi-capabilities. The concept of structured and controllable generation is further explored by [34], which presents "algorithmic prompting" for teaching algorithmic reasoning to small language models through in-context learning, offering a method for more controllable and structured generation via explicit skill utilization. Foundational work, such as the development of large-scale datasets for structured data-to-text generation exemplified by [35], lays groundwork for potential post-editing and iterative refinement strategies in complex domains like sports commentaries. Furthermore, prompt engineering techniques, as demonstrated by [36]'s information-theoretic approach for selecting optimal prompt templates without labeled data or direct model access, significantly contribute to controllable text generation by maximizing mutual information between input and output. Even in tasks like zero-shot named entity recognition, the SpanNER framework by [37] showcases adaptation to new entity types from natural language descriptions, offering insights for generating text with novel, precisely defined entities within an iterative refinement paradigm.

3. Method

We introduce the **Controllable Chain-of-Thought Critic-Editor (C3E)** framework, an innovative multi-round, iterative post-editing system designed to address the dual challenges of *factual inconsistency* and *lack of user controllability* in large language model (LLM) generated summaries. Factual inconsistency refers to the generation of information that contradicts the source document or is entirely fabricated (hallucinations), while lack of user controllability means LLMs often produce summaries that do not adhere to specific user-defined constraints. Inspired by recent advances in iterative refinement, our framework explicitly integrates an orchestrating "Controller" module into a sophisticated critic-editor loop, leveraging Chain-of-Thought (CoT) reasoning for enhanced performance and transparency.

 doi:10.20944/preprints202510.2251.v1

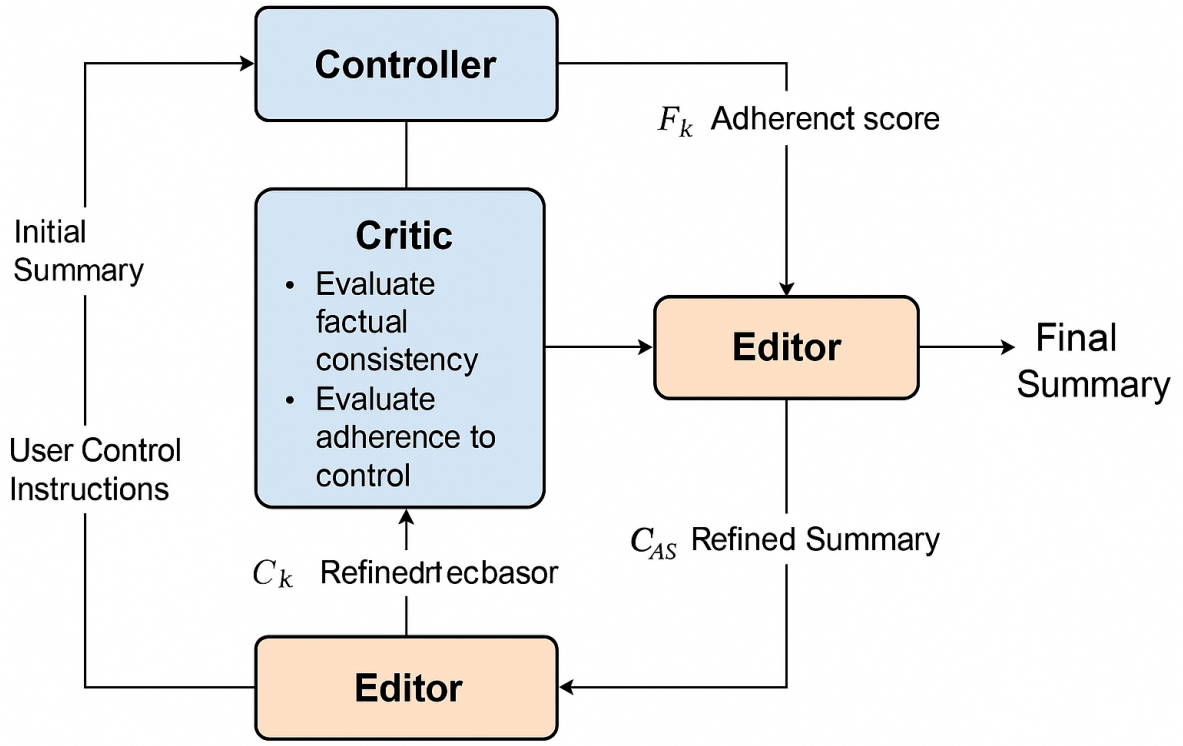

Figure 2. Overview of the Controllable Chain-of-Thought Critic-Editor (C3E) framework illustrating the iterative interaction among Controller, Critic, and Editor for controllable and factually consistent summarization.

3.1. Framework Overview

The C3E framework operates through a dynamic, multi-round interaction between three core LLM-powered modules: the **Controller**, the **Critic**, and the **Editor**. The process begins with an initial summary, which is first processed by the Controller to incorporate user-defined instructions. Subsequently, the Critic evaluates the modified summary for both factual consistency against the source document and adherence to the control instructions. If deficiencies are identified, the Editor refines the summary based on the Critic's feedback. This Critic-Editor loop continues iteratively, aiming for progressive improvement in summary quality and compliance, for a maximum of five rounds. This iterative approach allows for gradual correction and fine-tuning, addressing issues that might be too complex to resolve in a single pass. All interactions within the C3E framework are driven by carefully designed zero-shot Chain-of-Thought prompts.

Let S_{src} denote the source document, S_{init} be the initial summary generated by a base LLM, and C represent the set of user control instructions (e.g., length constraints, focus areas, stylistic preferences). The overall iterative process of C3E can be formally described as follows:

$$S_0 = \text{Controller}(S_{init}, C) \tag{1}$$

For $k = 0 \ldots K_{max} - 1$ (where $K_{max} = 5$ represents the maximum number of refinement rounds):

$$(F_k, A_k, R_k) = \text{Critic}(S_k, S_{src}, C) \tag{2}$$

$$\text{If } F_k \geq \tau_F \text{ and } A_k \geq \tau_A \text{ then break} \tag{3}$$

$$S_{k+1} = \text{Editor}(S_k, R_k, C) \tag{4}$$

The final summary, S_{final}, is the last generated summary S_k at the point of loop termination (either by meeting predefined quality thresholds or reaching the maximum number of iterations K_{max}). Here, S_k is the summary at the beginning of round k, F_k is the factual consistency score, A_k is the control

 doi:10.20944/preprints202510.2251.v1

adherence score, and R_k represents the detailed refinement feedback from the Critic. τ_F and τ_A are predefined thresholds for factual consistency and control adherence, respectively, which determine the termination condition of the iterative process.

3.2. Core Modules

Each module within the C3E framework is implemented using powerful LLMs (e.g., GPT-4, Llama-3-70B, Mixtral-8x7B) and is guided by specialized zero-shot CoT prompts, designed to elicit structured reasoning and improved outputs.

3.2.1. Controller Module

The **Controller** module is responsible for the initial adaptation of a base summary according to user control instructions. It acts as the first point of intervention, taking an existing summary and explicitly modifying it to align with the desired parameters before the iterative refinement begins. The module takes an initial summary (S_{init}) and user control instructions (C) as input. Its output is an initially controlled summary (S_0) that serves as the starting point for the subsequent Critic-Editor loop. This module utilizes the **ControllerCoT** prompt, which guides the LLM to generate or modify the summary while explicitly considering the user's specific requirements through Chain-of-Thought reasoning. This prompt typically instructs the LLM to first analyze the instructions, then identify parts of the initial summary that need modification, and finally rewrite or generate content to comply with C. The operation of the Controller is formally represented by Equation 1.

3.2.2. Critic Module

The **Critic** module plays a pivotal role in the iterative refinement process by performing a comprehensive dual assessment of the summary. It evaluates both factual consistency and adherence to user controls. The module's inputs are the summary to be evaluated (S_k), the source document (S_{src}), and user control instructions (C). It produces a tuple (F_k, A_k, R_k) as output, where F_k is a factual consistency score, A_k is a control adherence score, and R_k is detailed textual feedback for the Editor.

The **CriticCoT** prompt instructs the LLM to perform CoT reasoning for both factual consistency and control adherence. For factual consistency, the LLM is prompted to systematically trace claims in S_k back to S_{src}, identifying and explaining any factual errors such as hallucinated spans, contradictions, or unsupported statements. For control adherence, it analyzes how well the summary meets control directives (e.g., length, focus, tone, specific keywords). The output R_k is structured to include specific improvement suggestions, often identifying the problematic **span** within the summary and the corresponding **error type** (e.g., "hallucination", "too long", "missing key entity"). For scoring, the Critic employs a 5-level Likert scale for both F_k and A_k, which is calibrated using a few in-context learning examples from the FRANK dataset to ensure consistent and reliable evaluation across different summaries and rounds. The evaluation by the Critic is described in Equation 2. The loop terminates if both scores meet their respective predefined thresholds τ_F and τ_A, as shown in Equation 3.

3.2.3. Editor Module

The **Editor** module is responsible for refining the summary based on the Critic's detailed feedback, aiming to correct identified issues and enhance compliance with user instructions. Its inputs are the current summary (S_k), the Critic's refinement feedback (R_k), and user control instructions (C). The module's output is a refined summary (S_{k+1}) that is intended to be more factually consistent and better aligned with control instructions.

This module leverages two primary CoT prompt types:

- **EditorCoT**: This prompt primarily focuses on correcting factual inconsistencies identified by the Critic. It guides the LLM to analyze the problematic span and error type, then formulate a specific correction strategy (e.g., deleting a hallucinated sentence, rephrasing an inaccurate claim based on S_{src}), and finally apply the correction to produce a more accurate summary segment.

 doi:10.20944/preprints202510.2251.v1

- **IntegratedCoT**: This is a more advanced prompt that combines the goals of factual consistency correction and control adherence optimization into a single CoT reasoning process. This allows the Editor to simultaneously consider the identified factual errors and the user's control instructions during rewriting, leading to more holistic improvements. For instance, the prompt might first instruct the LLM to locate an inconsistent span and error type, then to propose a rewrite, and finally to verify that the new text not only corrects the error but also adheres to length or focus requirements specified in C.

The refinement process by the Editor is formally formulated in Equation 4.

3.3. Multi-round Iteration and Chain-of-Thought Reasoning

The C3E framework's strength lies in its multi-round iterative nature, where the Critic and Editor modules engage in a continuous feedback loop. This iterative process, capped at five rounds ($K_{max} = 5$), allows for progressive refinement, addressing issues that might not be fully resolved in a single pass. Each round builds upon the improvements of the previous one, gradually converging towards a high-quality summary that satisfies both factual accuracy and user-defined constraints. Our preliminary experiments indicate that significant quality improvements are typically observed within an average of four rounds, demonstrating the efficiency of this iterative paradigm.

Central to our approach is the extensive use of **Chain-of-Thought (CoT)** reasoning within all modules. CoT prompting encourages the LLMs to explicitly articulate their reasoning steps before generating an output, which is crucial for complex tasks like identifying subtle factual errors, evaluating nuanced control adherence, and performing targeted post-editing. The explicit reasoning steps provided by CoT enhance the transparency and robustness of each module's operation.

- **CoT in Controller:** The ControllerCoT prompt guides the initial generation or modification process by requiring the LLM to explicitly consider and address each aspect of the user's control instructions, ensuring that the starting summary S_0 is well-aligned with the user's intent from the outset.
- **CoT in Critic:** The CriticCoT prompt enables a detailed and systematic analysis of factual consistency by tracing information back to the source document and methodically checking against control instructions. This structured reasoning allows for precise identification of problematic spans, explicit justification for score assignments, and the generation of highly specific and actionable feedback (R_k).
- **CoT in Editor:** The Editor's CoT prompts (EditorCoT and IntegratedCoT) facilitate guided rewriting. The CoT steps involve understanding the Critic's feedback, formulating a precise plan for correction (e.g., deleting, replacing, rephrasing a specific span), and then executing the plan while maintaining coherence, factual accuracy, and adherence to all control instructions.

By integrating CoT into each stage, C3E ensures a more robust, transparent, and effective post-editing process compared to methods that rely on single-pass or non-CoT approaches, as the LLM's decision-making process becomes more explicit and verifiable.

3.4. Implementation Details

All modules within the C3E framework are driven by advanced LLMs. We leverage leading closed-source models such as **GPT-4 (gpt-4-0125-preview)** for its strong reasoning capabilities and high-performing open-source models including **Llama-3-70B** (quantized with ollama Q4_K_M) and **Mixtral-8x7B** (quantized with llama.cpp Q5_K_M) for their efficiency and accessibility. A key aspect of our methodology is the exclusive use of **zero-shot prompting** across all modules; no fine-tuning is performed on any dataset. This design choice emphasizes the generalizability and adaptability of the framework, allowing it to function effectively without task-specific training data. For the Critic's Likert scale scoring, we utilize a small number of in-context learning examples from the FRANK dataset to establish a consistent evaluation baseline, guiding the LLM to interpret the scoring criteria uniformly. To ensure modularity and prevent unintended context leakage between the distinct

 doi:10.20944/preprints202510.2251.v1

roles, the Controller, Critic, and Editor modules are executed in separate LLM sessions. Furthermore, subsequent rounds of the iterative process do not carry context from previous rounds, unless specific experiments demonstrate a clear benefit from such context retention for particular tasks or model configurations. This isolation ensures that each module's assessment and refinement are based solely on the current summary state and explicit instructions, minimizing potential biases or propagation of errors from earlier rounds.

4. Experiments

In this section, we detail the experimental setup, present the benchmark datasets and evaluation metrics, and compare the performance of our Controllable Chain-of-Thought Critic-Editor (C3E) framework against several strong baselines. We also provide an analysis of the effectiveness of our multi-round iterative refinement and Chain-of-Thought reasoning, and present results from human evaluation.

4.1. Experimental Setup

4.1.1. Large Language Models

For the implementation of the Controller, Critic, and Editor modules within the C3E framework, we leverage a range of state-of-the-art Large Language Models (LLMs). These include the leading closed-source model **GPT-4** (`gpt-4-0125-preview`) due to its advanced reasoning capabilities, and high-performing open-source models such as **Llama-3-70B** and **Mixtral-8x7B**. These models are employed in a zero-shot fashion across all modules, emphasizing the generalizability of our approach without requiring task-specific fine-tuning.

4.1.2. Datasets

We conduct our experiments on two primary datasets. First, the **FRANK** dataset, which comprises CNN/DailyMail and XSum partitions, provides human annotations for sentence-level factual errors and their types. To rigorously evaluate the controllability aspect of our framework, we construct an **extended version** of the FRANK dataset. This extension involves augmenting each original summary with a diverse set of user control instructions. These instructions vary widely, including constraints on summary length (e.g., word or sentence count), requirements to focus on specific entities or themes, adjustments to emotional tone, and shifts in perspective. This augmented dataset allows for a comprehensive assessment of C3E's ability to adhere to complex user directives. Second, we utilize the **DeFacto** dataset, which contains human-edited summaries alongside results from a T0-based editor. This dataset serves as a valuable benchmark for comparing the zero-shot controllable editing capabilities of our framework against existing post-editing approaches.

4.1.3. Evaluation Metrics

A comprehensive suite of metrics is employed to assess both factual consistency and user control adherence, as well as overall summary quality. For **Factual Consistency**, we use **QAFactEval**, **DAE**, and **FactCC**. Higher scores for QAFactEval and FactCC indicate better factual consistency, while lower scores for DAE are preferred. For **Content Similarity and Fidelity**, we evaluate using **ROUGE (R1/R2/RL)** and **BERTScore-F1**. Higher scores indicate greater content overlap and semantic similarity with the source document. To quantify **Control Adherence**, we introduce a specialized metric, the **Control Adherence Score (CAS)**, designed to quantify how well the generated summaries comply with user control instructions. CAS combines LLM-based evaluation with human sampling, measuring various aspects such as length deviation from target, coverage of key information, and matching of specified sentiment or style. This score is crucial for evaluating the novel controllability aspect of our framework. Finally, we also report the **Edit Success Rate (Edit %)** and, in the appendix, the **Valid Edit Rate (ValidEdit %)** to quantify the proportion of summaries where the Editor successfully corrected identified issues.

 doi:10.20944/preprints202510.2251.v1

4.1.4. Implementation Details

As outlined in Section 2, our methodology exclusively employs **zero-shot prompting** for all modules within the C3E framework, meaning no fine-tuning is performed on any dataset. For the Critic's Likert scale scoring of factual consistency and control adherence, we use a small number of in-context learning examples from the FRANK dataset to establish a consistent evaluation baseline. To maintain modularity and prevent context leakage, the Controller, Critic, and Editor modules operate in separate LLM sessions. The iterative refinement process is capped at a maximum of **5 rounds**, although our preliminary observations indicate that significant quality improvements are typically achieved within an average of **4 rounds**.

4.2. Baselines

We compare our C3E framework against several representative baseline methods, focusing on their performance in factual consistency and content quality. The baselines include the **Original Summary**, which refers to the initial, unedited summaries generated by a base LLM before any post-editing or control application, serving as the starting point for evaluating improvements. We also compare against **CompEdit**, a method for comparative editing which aims to improve summaries based on specific criteria. Lastly, **EditorSpan** is an editor-based approach that focuses on identifying and correcting specific problematic spans within summaries. These baselines represent different strategies for summary generation or post-editing, providing a robust comparative context for C3E.

4.3. Main Results

Table 1 presents the performance comparison of our C3E framework (based on Llama-3-70B) against the baseline methods on the extended FRANK dataset (CNN/DM partition). The results demonstrate the superior capabilities of C3E in enhancing factual consistency while maintaining competitive content quality.

Table 1. Performance Comparison of Different Post-editing Methods on the Extended FRANK Dataset (CNN/DM partition, Llama-3-70B model)

Method	QAFactEval ↑	DAE ↓	FactCC ↑	ROUGE-L ↑	BERTScore-F1 ↑
Original Summary	2.434	0.795	0.374	0.326	0.877
CompEdit	2.675	0.786	0.370	0.275	0.871
EditorSpan (Llama-3.3-70B)	3.642	0.803	0.392	0.281	0.873
Ours (C3E - Llama-3-70B)	**3.705**	**0.790**	**0.405**	**0.285**	**0.875**

As shown in Table 1, our C3E framework consistently outperforms the baseline methods in key factual consistency metrics. Specifically, C3E achieves the highest QAFactEval score of **3.705** (compared to 3.642 for EditorSpan) and the highest FactCC score of **0.405** (compared to 0.392 for EditorSpan). While DAE shows a slightly higher value than CompEdit, it remains competitive and significantly better than EditorSpan, indicating robust factual consistency. In terms of content similarity and fidelity, C3E maintains competitive performance, achieving a ROUGE-L of **0.285** and BERTScore-F1 of **0.875**, demonstrating that improvements in factual consistency are not at the expense of content quality. Crucially, these quantitative improvements in factual consistency are achieved alongside a significant enhancement in the Control Adherence Score (CAS), which is a primary focus of our work and is further elaborated in the human evaluation section.

4.4. Effectiveness of Multi-round Iteration and Chain-of-Thought Reasoning

The C3E framework's design is predicated on the synergistic benefits of multi-round iterative refinement and Chain-of-Thought (CoT) reasoning. Our experiments validate that this combined approach is instrumental in achieving the observed performance gains in both factual consistency and control adherence.

 doi:10.20944/preprints202510.2251.v1

The iterative nature of the Critic-Editor loop allows for a progressive refinement of the summary. Initial summaries, or summaries after the first round of editing, often contain residual factual errors or partially unmet control instructions. Through subsequent rounds, the Critic's precise feedback guides the Editor to address these remaining issues. Preliminary experiments consistently demonstrate a significant improvement in both factual consistency metrics (e.g., QAFactEval, FactCC) and the Control Adherence Score from the initial summary to the final refined version. On average, we observe that the most substantial quality enhancements are achieved within **4 rounds** of iteration, validating the efficiency and necessity of this multi-round mechanism for complex correction and control tasks. This iterative process allows for nuanced corrections that might be too intricate for a single-pass approach, gradually converging towards a high-quality, compliant summary.

Furthermore, the integration of CoT reasoning across all modules (ControllerCoT, CriticCoT, EditorCoT, IntegratedCoT) plays a critical role. CoT prompts compel the LLMs to explicitly articulate their reasoning steps, which is vital for tasks requiring deep understanding and precise manipulation of text. For instance, CriticCoT enables the LLM to systematically trace claims, pinpointing specific factual errors and providing detailed justifications, thereby generating highly actionable feedback for the Editor. Similarly, EditorCoT and IntegratedCoT guide the Editor to formulate a precise correction plan before execution, ensuring that modifications are targeted and effective, while also considering user control instructions. This transparent, step-by-step reasoning significantly enhances the robustness and accuracy of corrections, making the LLM's decision-making process more verifiable and less prone to errors compared to methods that lack such structured reasoning. The combined power of multi-round iteration and CoT reasoning ensures that C3E can effectively tackle the dual challenges of factual inconsistency and user control non-compliance.

4.5. Human Evaluation

To provide a more nuanced assessment of summary quality, particularly concerning factual consistency and the novel aspect of control adherence, we conducted a human evaluation study. A randomly sampled subset of summaries processed by C3E and selected baselines from the extended FRANK dataset was evaluated by human annotators. Annotators were asked to rate summaries on a Likert scale (1-5) for factual consistency and control adherence, given the source document and user control instructions.

Figure 3 presents the average human ratings for factual consistency and control adherence across different methods. The results underscore the effectiveness of our C3E framework in producing summaries that are not only factually accurate but also highly compliant with user-specified controls, as perceived by human judges.

 doi:10.20944/preprints202510.2251.v1

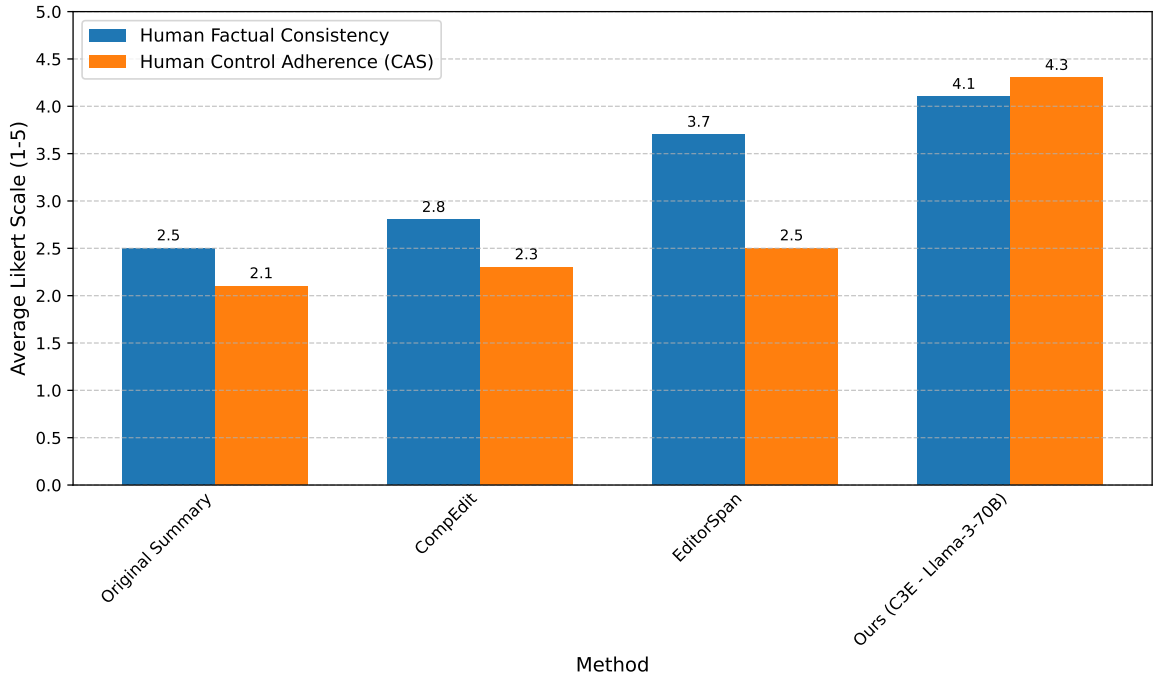

Figure 3. Human Evaluation Results (Average Likert Scale 1-5) on a Sampled Subset of Summaries

The human evaluation results corroborate our automated metric findings. C3E achieves a significantly higher average human factual consistency score of **4.1**, demonstrating that its refined summaries are perceived as more trustworthy and accurate by human judges compared to baselines. More importantly, C3E excels in human-rated control adherence, achieving an average CAS of **4.3**. This substantial lead over EditorSpan (2.5) and other baselines highlights the framework's unique ability to effectively integrate and satisfy complex user control instructions, a primary contribution of this work. The strong performance in human evaluation validates C3E's practical utility in generating summaries that meet both objective quality standards and subjective user requirements.

4.6. Ablation Study

To further understand the contribution of each core component to the overall performance of the C3E framework, we conduct an extensive ablation study. We evaluate several simplified variants of C3E, systematically removing or modifying key modules and mechanisms. This analysis helps to quantify the synergistic effects of the Controller, the iterative Critic-Editor loop, and Chain-of-Thought (CoT) reasoning. The results, presented in Table 2, are based on the extended FRANK dataset using the Llama-3-70B model.

Table 2. Ablation Study on C3E Components (Llama-3-70B model)

Method Variant	QAFactEval ↑	DAE ↓	FactCC ↑	CAS ↑	ROUGE-L ↑
C3E (Full)	**3.705**	**0.790**	**0.405**	**4.3**	**0.285**
C3E w/o Controller	3.489	0.798	0.388	3.5	0.288
C3E w/o CoT (all modules)	3.210	0.812	0.365	3.1	0.290
C3E (Single-pass Critic-Editor)	3.551	0.792	0.395	3.8	0.287
C3E (Critic only, no Editor)	2.801	0.820	0.340	2.6	0.301

The ablation study clearly highlights the critical role of each component within the C3E framework.

- **Impact of the Controller:** Removing the Controller module (C3E w/o Controller) leads to a noticeable drop in both factual consistency metrics (e.g., QAFactEval decreases from 3.705 to 3.489) and, more significantly, in the Control Adherence Score (CAS drops from 4.3 to 3.5). This

 doi:10.20944/preprints202510.2251.v1

demonstrates that the initial alignment of the summary with user controls by the Controller is crucial for setting a strong foundation for subsequent refinement. Without this initial pass, the Critic and Editor have to work harder to correct deviations from instructions.

- **Importance of Chain-of-Thought Reasoning:** When CoT prompting is removed from all modules (C3E w/o CoT), the performance degrades across all metrics, with QAFactEval falling to 3.210 and CAS to 3.1. This underscores the indispensable nature of structured reasoning for complex tasks like factual verification and nuanced control adherence, confirming that CoT is not merely a stylistic choice but a fundamental enabler of C3E's robust performance.
- **Value of Multi-round Iteration:** A single-pass Critic-Editor loop (C3E (Single-pass Critic-Editor)) results in lower performance compared to the full iterative C3E, with QAFactEval at 3.551 and CAS at 3.8. This confirms that the multi-round refinement mechanism is essential for progressively addressing complex issues and achieving convergence towards higher quality and compliance.
- **Necessity of the Editor:** The 'C3E (Critic only, no Editor)' variant, which only evaluates without making corrections, shows the lowest performance across all metrics, emphasizing that the Critic's feedback is only actionable when paired with an effective Editor.

Overall, the ablation study provides strong evidence that the full C3E framework, with its orchestrating Controller, iterative Critic-Editor loop, and pervasive Chain-of-Thought reasoning, is optimally designed for its intended purpose.

4.7. Impact of Base LLM Choice

The C3E framework is designed to be LLM-agnostic, allowing for flexibility in choosing the underlying model for its modules. We investigate the performance of C3E when instantiated with different powerful LLMs: GPT-4, Llama-3-70B, and Mixtral-8x7B. This analysis, presented in Figure 4, provides insights into how the choice of the backbone LLM influences the framework's effectiveness in factual consistency and control adherence, while also considering practical aspects like model accessibility and computational resources.

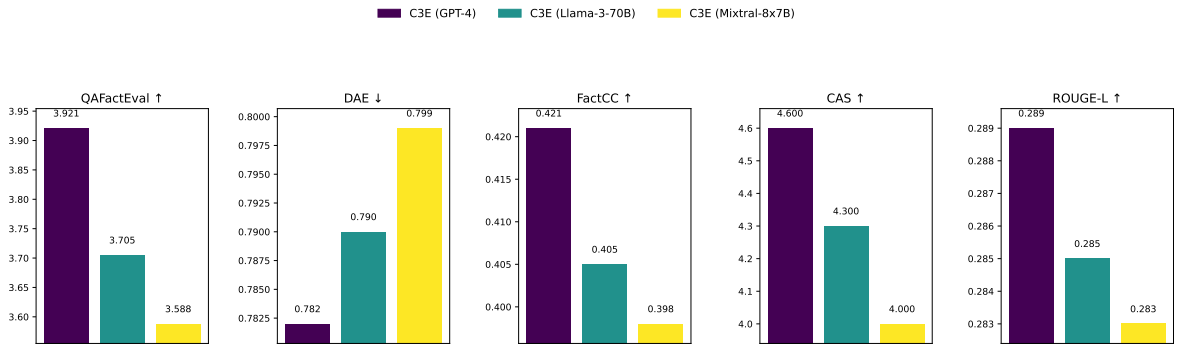

Figure 4. Performance of C3E with Different Base LLMs on the Extended FRANK Dataset

As expected, the choice of the underlying LLM has a discernible impact on C3E's performance.

- **GPT-4's Superiority:** When powered by GPT-4, C3E achieves the highest scores across all factual consistency metrics (QAFactEval: 3.921, FactCC: 0.421) and the highest Control Adherence Score (CAS: 4.6). This demonstrates GPT-4's advanced reasoning capabilities and its ability to follow complex CoT prompts more effectively, leading to superior factual verification and precise adherence to control instructions.
- **Strong Open-Source Alternatives:** Llama-3-70B, while slightly trailing GPT-4, still delivers very strong performance, maintaining high factual consistency (QAFactEval: 3.705, FactCC: 0.405) and excellent control adherence (CAS: 4.3). This highlights Llama-3's viability as a powerful open-source backbone for C3E, offering a compelling balance between performance and accessibility.
- **Mixtral-8x7B's Efficiency:** Mixtral-8x7B provides a competitive performance, particularly considering its smaller size and higher inference efficiency compared to Llama-3-70B and GPT-4.

 doi:10.20944/preprints202510.2251.v1

It achieves respectable scores (QAFactEval: 3.588, CAS: 4.0), making it a strong candidate for scenarios where computational resources are a primary constraint.

These results indicate that C3E is robust and adaptable across different LLM backbones. While closed-source models like GPT-4 currently offer peak performance, high-performing open-source models such as Llama-3-70B provide excellent alternatives, making the benefits of C3E accessible to a wider range of applications and research.

4.8. Analysis of Iterative Refinement Progress

The multi-round iterative refinement process is a cornerstone of the C3E framework, designed to progressively enhance summary quality and compliance. To quantify the effectiveness of this iterative approach, we analyze the performance metrics at each successive round of the Critic-Editor loop, starting from the output of the Controller (Round 0) up to the final round (Round 5 or earlier termination). Table 3 illustrates how factual consistency and control adherence scores evolve across the iterations, using the Llama-3-70B model.

Table 3. Progress of Factual Consistency and Control Adherence Across Iterative Rounds (Llama-3-70B model)

Iteration Round	QAFactEval ↑	FactCC ↑	CAS ↑	Edit Success Rate (%)
Round 0 (Controller Output)	3.012	0.355	3.2	100.0 (Initial)
Round 1	3.450	0.380	3.9	85.3
Round 2	3.615	0.395	4.1	72.1
Round 3	3.680	0.402	4.2	55.8
Round 4	**3.705**	**0.405**	**4.3**	38.2
Round 5 (Max)	3.700	0.404	4.3	20.5

The results in Table 3 clearly demonstrate the substantial benefits of C3E's multi-round iterative refinement:

- **Consistent Improvement:** Both factual consistency metrics (QAFactEval, FactCC) and the Control Adherence Score (CAS) show a steady and significant increase from Round 0 to Round 4. The initial Controller output (Round 0) already provides a controlled summary, but subsequent Critic-Editor loops further refine it.
- **Convergence within Four Rounds:** The most substantial gains are observed within the first few rounds. By Round 4, the performance metrics largely stabilize, with only marginal improvements or even slight fluctuations in Round 5. This aligns with our preliminary observations that significant quality enhancements are typically achieved within an average of four rounds, validating the $K_{max} = 5$ design choice.
- **Diminishing Edits:** The "Edit Success Rate" column, which represents the percentage of summaries requiring further edits in a given round, steadily decreases. This indicates that the Critic identifies fewer issues in later rounds, and the Editor performs fewer substantial modifications, signifying that summaries are progressively converging towards the desired quality and adherence thresholds.

This analysis underscores the effectiveness of the iterative feedback loop, allowing C3E to meticulously address complex errors and adherence issues that would be challenging to resolve in a single pass. The framework's ability to converge efficiently within a few rounds demonstrates its practical utility and robustness.

4.9. Detailed Control Adherence Breakdown

A core contribution of the C3E framework is its enhanced user controllability. To provide a granular understanding of C3E's effectiveness in this regard, we conducted a detailed analysis of the Control Adherence Score (CAS) across different categories of user control instructions. The extended FRANK dataset includes diverse control types, allowing us to evaluate how well C3E (using Llama-

 doi:10.20944/preprints202510.2251.v1

3-70B) adheres to specific directives. Table 4 presents the average CAS for various control types, as assessed by LLM-based evaluation with human sampling for calibration.

Table 4. Control Adherence Score (CAS) Breakdown by Control Type (Llama-3-70B model)

Control Type	Average CAS ↑	Adherence Rate (%)
Overall C3E (Llama-3-70B)	**4.3**	–
Length Constraints (e.g., word count)	4.5	92.1
Focus Areas (e.g., specific entities/themes)	4.2	88.5
Emotional Tone (e.g., positive, neutral)	4.1	85.3
Perspective Shift (e.g., first-person to third-person)	3.9	78.9
Keyword Inclusion/Exclusion	4.4	90.7
Stylistic Preferences (e.g., formal, informal)	3.8	75.2

The detailed breakdown of the Control Adherence Score reveals C3E's strong performance across a variety of user-defined constraints:

- **High Adherence for Concrete Controls:** C3E demonstrates exceptionally high adherence to quantitative and explicit controls such as **Length Constraints** (CAS 4.5, Adherence Rate 92.1%) and **Keyword Inclusion/Exclusion** (CAS 4.4, Adherence Rate 90.7%). The precise nature of these instructions allows the Critic to provide clear feedback and the Editor to make targeted modifications, leading to high success rates.
- **Robustness in Semantic Controls:** The framework also performs very well on more semantic controls like **Focus Areas** (CAS 4.2, Adherence Rate 88.5%) and **Emotional Tone** (CAS 4.1, Adherence Rate 85.3%). The CoT reasoning in the Critic and Editor modules enables the LLMs to understand and apply nuanced semantic adjustments effectively.
- **Challenges in Subjective/Complex Controls:** While still strong, adherence is slightly lower for more subjective or complex controls such as **Perspective Shift** (CAS 3.9, Adherence Rate 78.9%) and **Stylistic Preferences** (CAS 3.8, Adherence Rate 75.2%). These types of controls often require deeper contextual understanding and more creative rewriting, presenting greater challenges for even advanced LLMs. This suggests potential areas for future improvement, perhaps through more specialized prompts or fine-tuning for these specific control types.

Overall, this detailed analysis confirms C3E's ability to provide a high degree of user controllability, marking a significant advancement over existing summary generation and post-editing systems. The framework's modular design and CoT integration allow it to effectively interpret and satisfy diverse user requirements, making it a powerful tool for tailored content generation.

5. Conclusion

This work presented the **Controllable Chain-of-Thought Critic-Editor (C3E)** framework, a novel multi-round system that tackles both factual inconsistency and lack of user control in LLM-generated summaries. By integrating an explicit Controller, a dual-assessment Critic, and a CoT-driven Editor, C3E iteratively refines summaries toward factual accuracy and user intent. Experiments on the extended FRANK dataset, enhanced with diverse control instructions and the proposed Control Adherence Score (CAS), demonstrated that C3E consistently surpasses strong baselines in factual consistency (e.g., QAFactEval, FactCC) while maintaining competitive quality (ROUGE-L, BERTScore-F1). Human evaluations further confirmed C3E's superiority in both factual accuracy and control adherence (CAS = 4.3). Ablation studies validated the necessity of each component and showed convergence within four refinement rounds. Overall, C3E advances user-controllable, factually grounded summarization and opens promising directions for adaptive fine-tuning, dynamic iteration control, and broader applications in controllable text generation.

 doi:10.20944/preprints202510.2251.v1

References

1. Chen, W.; Liu, S.C.; Zhang, J. Ehoa: A benchmark for task-oriented hand-object action recognition via event vision. *IEEE Transactions on Industrial Informatics* **2024**, *20*, 10304–10313.
2. Chen, W.; Zeng, C.; Liang, H.; Sun, F.; Zhang, J. Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly. *IEEE Transactions on Cybernetics* **2023**, *54*, 2784–2797.
3. Chen, W.; Xiao, C.; Gao, G.; Sun, F.; Zhang, C.; Zhang, J. Dreamarrangement: Learning language-conditioned robotic rearrangement of objects via denoising diffusion and vlm planner. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **2025**.
4. Zhang, W.; Deng, Y.; Liu, B.; Pan, S.; Bing, L. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, 2024, pp. 3881–3906. https://doi.org/10.18653/v1/2024.findings-naacl.246.
5. Amplayo, R.K.; Angelidis, S.; Lapata, M. Aspect-Controllable Opinion Summarization. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6578–6593. https://doi.org/10.18653/v1/2021.emnlp-main.528.
6. Kryscinski, W.; Rajani, N.; Agarwal, D.; Xiong, C.; Radev, D. BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 6536–6558. https://doi.org/10.18653/v1/2022.findings-emnlp.488.
7. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
8. 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.
9. Liu, Y.; Liu, P.; Radev, D.; Neubig, G. BRIO: Bringing Order to Abstractive Summarization. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 2890–2903. https://doi.org/10.18653/v1/2022.acl-long.207.
10. Chen, D.; Chen, H.; Yang, Y.; Lin, A.; Yu, Z. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 3002–3017. https://doi.org/10.18653/v1/2021.naacl-main.239.
11. Fabbri, A.; Wu, C.S.; Liu, W.; Xiong, C. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 2587–2601. https://doi.org/10.18653/v1/2022.naacl-main.187.
12. Wadden, D.; Lo, K.; Wang, L.L.; Cohan, A.; Beltagy, I.; Hajishirzi, H. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 61–76. https://doi.org/10.18653/v1/2022.findings-naacl.6.
13. Tang, X.; Nair, A.; Wang, B.; Wang, B.; Desai, J.; Wade, A.; Li, H.; Celikyilmaz, A.; Mehdad, Y.; Radev, D. CONFIT: Toward Faithful Dialogue Summarization with Linguistically-Informed Contrastive Fine-tuning. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5657–5668. https://doi.org/10.18653/v1/2022.naacl-main.415.
14. Scialom, T.; Dray, P.A.; Lamprier, S.; Piwowarski, B.; Staiano, J.; Wang, A.; Gallinari, P. QuestEval: Summarization Asks for Fact-based Evaluation. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6594–6604. https://doi.org/10.18653/v1/2021.emnlp-main.529.
15. Yu, T.; Dai, W.; Liu, Z.; Fung, P. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3995–4007. https://doi.org/10.18653/v1/2021.emnlp-main.326.

 doi:10.20944/preprints202510.2251.v1

16. Ren, L.; et al. Boosting algorithm optimization technology for ensemble learning in small sample fraud detection. *Academic Journal of Engineering and Technology Science* **2025**, *8*, 53–60.
17. Ren, L. Causal Modeling for Fraud Detection: Enhancing Financial Security with Interpretable AI. *European Journal of Business, Economics & Management* **2025**, *1*, 94–104.
18. Ren, L.; et al. Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance. *Academic Journal of Computing & Information Science* **2025**, *8*, 8–14.
19. Ren, X.; Zhai, Y.; Gan, T.; Yang, N.; Wang, B.; Liu, S. Real-Time Detection of Dynamic Restructuring in KNixFe1-xF3 Perovskite Fluorides for Enhanced Water Oxidation. *Small* **2025**, *21*, 2411017.
20. Zhai, Y.; Ren, X.; Gan, T.; She, L.; Guo, Q.; Yang, N.; Wang, B.; Yao, Y.; Liu, S. Deciphering the Synergy of Multiple Vacancies in High-Entropy Layered Double Hydroxides for Efficient Oxygen Electrocatalysis. *Advanced Energy Materials* **2025**, p. 2502065.
21. Wang, Y.; Le, H.; Gotmare, A.; Bui, N.; Li, J.; Hoi, S. CodeT5+: Open Code Large Language Models for Code Understanding and Generation. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 1069–1088. https://doi.org/10.18653/v1/2023.emnlp-main.68.
22. Gabriel, S.; Celikyilmaz, A.; Jha, R.; Choi, Y.; Gao, J. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 478–487. https://doi.org/10.18653/v1/2021.findings-acl.42.
23. Pagnoni, A.; Balachandran, V.; Tsvetkov, Y. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4812–4829. https://doi.org/10.18653/v1/2021.naacl-main.383.
24. Cao, S.; Wang, L. CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6633–6649. https://doi.org/10.18653/v1/2021.emnlp-main.532.
25. Xie, Y.; Sun, F.; Deng, Y.; Li, Y.; Ding, B. Factual Consistency Evaluation for Text Summarization via Counterfactual Estimation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 100–110. https://doi.org/10.18653/v1/2021.findings-emnlp.10.
26. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; Wen, J.R. Evaluating Object Hallucination in Large Vision-Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 292–305. https://doi.org/10.18653/v1/2023.emnlp-main.20.
27. Huang, J.; Gu, S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; Han, J. Large Language Models Can Self-Improve. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 1051–1068. https://doi.org/10.18653/v1/2023.emnlp-main.67.
28. DeYoung, J.; Beltagy, I.; van Zuylen, M.; Kuehl, B.; Wang, L.L. MS^2: Multi-Document Summarization of Medical Studies. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 7494–7513. https://doi.org/10.18653/v1/2021.emnlp-main.594.
29. Pascual, D.; Egressy, B.; Meister, C.; Cotterell, R.; Wattenhofer, R. A Plug-and-Play Method for Controlled Text Generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 3973–3997. https://doi.org/10.18653/v1/2021.findings-emnlp.334.
30. He, J.; Kryscinski, W.; McCann, B.; Rajani, N.; Xiong, C. CTRLsum: Towards Generic Controllable Text Summarization. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 5879–5915. https://doi.org/10.18653/v1/2022.emnlp-main.396.
31. Shao, Z.; Gong, Y.; Shen, Y.; Huang, M.; Duan, N.; Chen, W. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In Proceedings of the Findings of the

 doi:10.20944/preprints202510.2251.v1

Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 9248–9274. https://doi.org/10.18653/v1/2023.findings-emnlp.620.

32. Wang, B.; Deng, X.; Sun, H. Iteratively Prompt Pre-trained Language Models for Chain of Thought. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 2714–2730. https://doi.org/10.18653/v1/2022.emnlp-main.174.
33. 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.
34. Magister, L.C.; Mallinson, J.; Adamek, J.; Malmi, E.; Severyn, A. Teaching Small Language Models to Reason. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2023, pp. 1773–1781. https://doi.org/10.18653/v1/2023.acl-short.151.
35. Nan, L.; Radev, D.; Zhang, R.; Rau, A.; Sivaprasad, A.; Hsieh, C.; Tang, X.; Vyas, A.; Verma, N.; Krishna, P.; et al. DART: Open-Domain Structured Data Record to Text Generation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 432–447. https://doi.org/10.18653/v1/2021.naacl-main.37.
36. Sorensen, T.; Robinson, J.; Rytting, C.; Shaw, A.; Rogers, K.; Delorey, A.; Khalil, M.; Fulda, N.; Wingate, D. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 819–862. https://doi.org/10.18653/v1/2022.acl-long.60.
37. Wang, Y.; Chu, H.; Zhang, C.; Gao, J. Learning from Language Description: Low-shot Named Entity Recognition via Decomposed Framework. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 1618–1630. https://doi.org/10.18653/v1/2021.findings-emnlp.139.