Divide-Conquer-Reasoning for Consistency Evaluation and Automatic Improvement of Large Language Models

Wendi Cui[1], **Zhuohang Li**[3], **Damien Lopez**[1], **Kamalika Das**[1,2], **Bradley Malin**[3,4], **Sricharan Kumar**[1,2], **Jiaxin Zhang**[1,2]*

[1]Intuit [2]Intuit AI Research [3]Vanderbilt University [4]Vanderbilt University Medical Center

{wendi_cui, jiaxin_zhang}@intuit.com, zhuohang.li@vanderbilt.edu

Abstract

Evaluating the quality and consistency of text generated by Large Language Models (LLMs) poses a significant, yet unresolved challenge for industry research. We propose DCR, an automated framework for evaluating and improving the consistency of LLM-generated texts using a divide-conquer-reasoning approach. Unlike existing LLM-based evaluators operating at the paragraph level, our method employs a divide-and-conquer evaluator (DCE) that breaks down the paragraph-to-paragraph comparison into sentence-to-paragraph comparisons. To facilitate this approach, we also introduce an automatic metric converter (AMC) that translates the output from DCE into an interpretable numeric score. Beyond the consistency evaluation, we further present a reason-assisted improver (RAI) that mitigates inconsistencies by leveraging the analytical reasons identified by DCE. Through comprehensive and systematic empirical analysis, we show that our approach outperforms state-of-the-art methods by a large margin (e.g., **+16.8**% and **+32.5**% on the SummEval dataset) in consistency evaluation across multiple benchmarks. Our approach also substantially reduces nearly 90% output inconsistencies in one iteration, showing promise for effective hallucination mitigation in real-world industrial applications.

1 Introduction

Large language models (LLMs) such as GPT-4 and PaLM 2 (Yang et al., 2023; Bubeck et al., 2023) have demonstrated impressive performance on a variety of natural language generation (NLG) tasks, including summarization (Tam et al., 2022), open-book question-answering (QA) (Kamalloo et al., 2023), and retrieval-augmented generation (RAG) (Lewis et al., 2020; Liu et al., 2023a). The evaluation of generated response quality often involves the assessment of the semantic equivalence between two pieces of text, e.g., between the generated response and the original text in summarization tasks or between two candidate responses in open-book QA tasks. However, conventional evaluation methods, such as BARTScore (Yuan et al., 2021) and BERTScore (Zhang et al., 2020), which rely on *token-level* comparison, are inadequate for accurately and reliably measuring the quality of generated content, particularly in complex scenarios with long paragraphs (Liu et al., 2023b; Hanna and Bojar, 2021). To address this issue, LLM-based evaluators such as G-Eval (Liu et al., 2023b) and GPTScore (Jinlan et al., 2023) have proposed a new framework that evaluates texts via *paragraph-level* comparison. While these evaluators show promise for certain tasks, their scores often fail to achieve high concordance with human judgments of semantic equivalence. Furthermore, as only numeric scores are provided with no explanation, it can be challenging for humans to trust or reason about these scores, particularly when using LLMs that are known to hallucinate (Li et al., 2023; Ji et al., 2023; Rawte et al., 2023).

Assessing the consistency of LLMs is more broadly connected to AI safety and has become a critical step in improving the reliability of these systems by preventing the generation of misinformation and harmful content. Wang et al. (2022) demonstrates that *consistency checking* can significantly enhance the chain of thought reasoning in LLMs. Similarly, Kuhn et al. (2023) leverages semantic consistency for uncertainty estimation in NLG. Recent studies employ consistency checking to detect hallucinations based on pre-trained LLMs (Manakul et al., 2023; Zhang et al., 2023a) and instruction-tuned LLMs (Mündler et al., 2023). Although these methods exhibit promising results on several specific tasks, including mathematical reasoning and factual assessment, the potential failures (Chen et al., 2023) of self-consistency are often overlooked. This is essentially due to a lack

*Corresponding Author. Our code is available at https://github.com/intuit-ai-research/DCR-consistency.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 334–361
November 12-16, 2024 ©2024 Association for Computational Linguistics

of a generic, automatic, and reliable strategy that assesses the consistency of two responses, let alone remediating such inconsistency.

In this work, we introduce a novel framework, `Divide-Conquer-Reasoning` (abbreviated as `DCR`), for developing an automatic and reliable consistency evaluation method. Our approach capitalizes on the intuition that human evaluators assess consistency by comparing the generated text to the reference text sentence by sentence and then combining the analysis to make a holistic judgment. Unlike existing metrics that rely on either token-level or paragraph-level checks, our approach breaks down the paragraph-to-paragraph comparison into a series of sentence-to-paragraph comparisons. This approach avoids confusing LLM by either providing too much information at once or zooming in too narrowly. Additionally, our approach does not rely on LLMs to directly output verbal scores in a regression manner, which have been shown to be prone to hallucination. We note that `DCR` is a reference-free method, which does not rely on a golden reference written by the human expert. For example in a summary task, `DCR` does not need a sample summary and can compare directly between the target summary and the original paragraphs.

2 Preliminaries

Black-Box LLM Evaluation. One of the drawbacks of current grey-box LLM evaluations is that they require output token-level probabilities (Jiang et al., 2023). However, prominent LLMs such as GPT-3.5, GPT-4, PaLM 2, and Claude 2, are only available through restricted API calls. Therefore, such token-level information might not be available. By contrast, we focus on the design of a black-box approach that remains applicable even when only text-based responses are available from the LLM; that is, we only have access to the model output.

Limitation of Existing Methods. The conventional metrics, such as BERTscore and BARTscore, rely on a *token-level* comparison using n-gram or contextual embedding to calculate cosine similarity. However, this approach fails to capture the overall semantic meaning as it directly aggregates token-level similarities. To address this issue, leveraging the power of LLMs for self-evaluation has been proposed. G-Eval (Liu et al., 2023b) and GPT-Eval (Jiang et al., 2023) evaluate consistency at a paragraph level by prompting LLMs to compare two candidates as a whole. However, these approaches have a major drawback as the generated verbal scores in a regression manner by LLMs are *prone to hallucinations*, resulting in abnormally higher ratings for LLM-generated content that diverge from human judgment (Liu et al., 2023b). Such methods also generate no actionable insight to justify the score or mitigate inconsistencies.

3 `DCR` Framework

To overcome the aforementioned limitations, we propose a Divide-Conquer-Reasoning framework, which comprises three essential components: (1) `DCE` disassembles the candidate paragraph, scrutinizes semantic inconsistencies using sentence-to-paragraph comparison and outputs sentence-level inconsistency/consistency reasons, (2) `AMC` converts such reasons into numeric scores for quantitative interpretation, and (3) `RAI` conducts analytical reasoning to improve consistency through candidate regeneration. As illustrated in Fig. 1, `DCR` involves a combination of sentence-level analysis, semantic consistency checking, and causal analysis, making it an ideal pipeline for a diverse range of tasks such as summarization, question-answering (QA), and retrieval-augmented generation (RAG). Moreover, `DCR` also improves the consistency of generated text through analysis and reasoning. Fig.2 provides an example of how `DCR` evaluates and enhances the consistency of the candidate text.

3.1 Divide-Conquer Evaluator (`DCE`)

The Divide-Conquer Evaluator (`DCE`) is an LLM Agent designed to perform semantic consistency checks using a sentence-to-paragraph strategy. It accepts a reference paragraph and a candidate paragraph as inputs. The reference paragraph does not need to be the ground truth or sample answer. For example in a summary task, the reference can be the original articles to be summarized. `DCE` breaks down the candidate paragraph into sentences (*divide*) and then assess each sentence against the reference (*conquer*). Given the input reference $\mathcal{R} = \langle s_1^r, ..., s_l^r \rangle$ and candidate $\mathcal{C} = \langle s_1^c, ..., s_k^c \rangle$, we build a `DCE` agent $\mathcal{L}_{\texttt{DCE}}$ using the LLM model $\mathcal{M}$ (e.g., GPT-3.5/4) with an instructed prompt $\mathcal{P}_{\texttt{DCE}}$ following Eq. (1):

$$\{\gamma_1, ..., \gamma_k\} = \mathcal{L}_{\texttt{DCE}}(\langle s_1^c, ..., s_k^c \rangle, \mathcal{R} \mid \mathcal{P}_{\texttt{DCE}}). \quad (1)$$

Eq. (1) generates *reasons*, denoted as $\Gamma = \{\gamma_1, ..., \gamma_k\}$, which is a list of reasons explaining

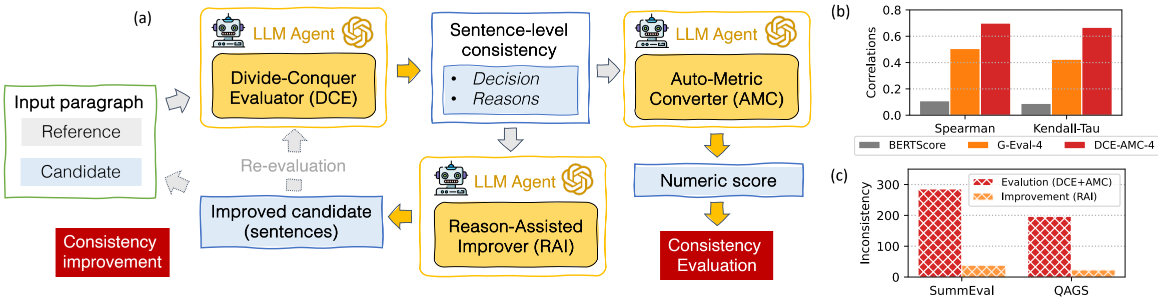

Figure 1: (a) Overview of the proposed DCR framework. The first two components (DCE-AMC) provide a better strategy for evaluating and quantifying semantic consistency to best match human judgments. Building on this, a third component RAI further utilizes analytical reasoning to iteratively mitigate spotted inconsistency in LLM-generated content w.r.t. the reference to mitigate hallucinations. (b) The combination of DCE and AMC significantly outperforms the baseline methods in terms of correlations with human ratings. (c) RAI substantially reduces output inconsistencies by $\sim 90\%$ through a single iteration on SummEval and QAGS.

why each sentence $s_i^c (i = 1, ..., k)$ is or is not consistent against the *entire* reference paragraph $\mathcal{R}$. We can tailor instruction prompts by defining task-specific criteria to accommodate different tasks. Table 6 provides an example prompt for the summarization consistency task. Since the comparison in DCE is not to a pair-wise comparison between sentences in the candidate text and that from the reference text *(sentence-to-sentence)*, but to compare each sentence in the candidate text sequence to the *entire* reference text sequence *(sentence-to-paragraph)*, it reduces the number of comparison operations and does not rely on any sentence-matching techniques, making it perfect to cover cases with a varying number of sentences (Amplayo et al., 2022).

3.2 Auto-Metric Converter (AMC)

The Auto-Metric Converter (AMC) is an LLM Agent that aims to quantitatively measure the consistency evaluation derived from the Divide-Conquer Evaluator (DCE) by converting the reasons from DCE into a numeric score system. This is accomplished by introducing an LLM agent, denoted as $\mathcal{L}_{\text{AMC}}$, which takes reasons $\langle \gamma_1, ..., \gamma_k \rangle$ with an instructed prompt $\mathcal{P}_{\text{AMC}}$ as inputs:

$$\{z_1, ..., z_k\} = \mathcal{L}_{\text{AMC}}(\{\gamma_1, ..., \gamma_k\} \mid \mathcal{P}_{\text{AMC}}). \quad (2)$$

The LLM Agent $\mathcal{L}_{\text{AMC}}$ functions as a binary sentiment classifier that classifies the reasons $\langle \gamma_1, ..., \gamma_k \rangle$ to be either positive (marked by "+1" if the sentence is consistent), or negative (marked by "-1" otherwise). As a result, AMC outputs an array of scores $\{z_1, ..., z_k\}, z_i \in \{-1, +1\}$ for each sentence $\langle s_1^c, ..., s_k^c \rangle$ in the candidate $\mathcal{C}$. We then utilize this score array to calculate a comprehensive score $\mathcal{Z}$ to evaluate how consistent the candidate (paragraph) is against the reference (paragraph):

$$\mathcal{Z} = \left(\sum_{i=1}^{k} z_i + \alpha \right) / (k + \beta), \ \hat{\mathcal{Z}} = \frac{(\mathcal{Z} + 1)}{2}, \quad (3)$$

where k is the length of the score array, i.e., the number of sentences in the candidate paragraph. Depending on the prompt, the *reasons* output by DCE may not all be on the sentence level. To ensure that the score calculated is solely generated by sentence-level *reasons*, we introduce α and β in Eq. (3), as explained in detail in Appendix I. Finally, we rescale $\mathcal{Z}$ to obtain the final score $\hat{\mathcal{Z}}$ to be between 0 (*completely inconsistent*) and 1 (*completely consistent*). A smaller $\hat{\mathcal{Z}}$ value indicates higher inconsistency between $\mathcal{C}$ and $\mathcal{R}$.

The AMC component serves as a binary sentiment classifier that classifies the reasons output by DCE to be either positive or negative for each sentence. It then utilizes such classifications to calculate a comprehensive score to evaluate consistency in a regression manner. Such a numerical score calculated by AMC is more stable than the verbal score directly output by LLMs. This design deliberately excludes the use of LLM in crucial steps where it tends to hallucinate or be biased, such as generating numerical evaluation scores, and harnesses its power where it has demonstrated excellence, such as classification and reasoning.

3.3 Reason-Assisted Improver (RAI)

The Reason-Assisted Improver (RAI) is an LLM Agent that focuses on improving the consistency of candidates by reasoning through the inconsistent explanations generated by the Divide-Conquer Evaluator (DCE). To achieve this goal, we propose

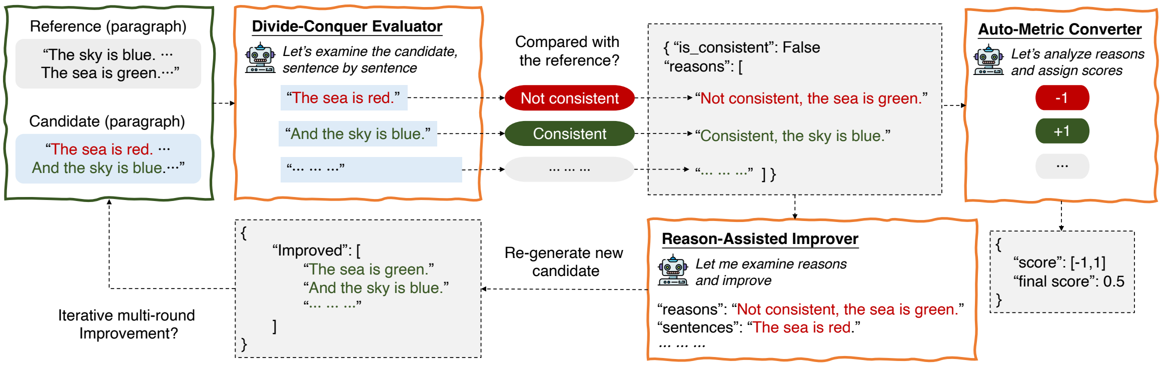

Figure 2: An example of evaluating and improving consistency via our proposed DCRframework.

an LLM agent $\mathcal{L}_{\texttt{RAI}}$ to generate new candidate sentences $\langle \hat{s}^c_1, ..., \hat{s}^c_k \rangle$ based on the collected reasons $\{\gamma_1, ..., \gamma_k\}$ and original sentences $\langle s^c_1, ..., s^c_k \rangle$:

$$\langle \hat{s}^c_1, ..., \hat{s}^c_k \rangle = \mathcal{L}_{\texttt{RAI}}(\gamma_1, ..., \gamma_k, \langle s^c_1, ..., s^c_k \rangle, \mathcal{R} | \mathcal{P}_{\texttt{RAI}}). \quad (4)$$

The core task of $\mathcal{L}_{\texttt{RAI}}$ is to rewrite the original sentence s^c_i if s^c_i is inconsistent with the reference $\mathcal{R}$ and return a new generated $\hat{s}^c_i$ ($\hat{s}^c_i \neq s^c_i$), otherwise retain s^c_i. The newly generated responses $\hat{\mathcal{C}} = \langle \hat{s}^c_1, ..., \hat{s}^c_k \rangle$ can either be returned as the improved answer or directly fed to the DCE agent in Eq. (1) to conduct *another-round* DCR , i.e., DCE→ AMC→ RAI, namely performing a *multi-round* consistency improvement, where the consistency is iteratively improved until reaching the maximum number of rounds m. Algorithm 1 illustrates the workflow of the DCR framework, which consists of three core components: DCE, AMC, and RAI.

Algorithm 1 Proposed DCR framework

Requirements: Candidate $\mathcal{C}$, Reference $\mathcal{R}$, LLM model $\mathcal{M}$, LLM agents $\mathcal{L}_{\texttt{DCE}}$, $\mathcal{L}_{\texttt{AMC}}$, $\mathcal{L}_{\texttt{RAI}}$ with instructed prompts $\mathcal{P}_{\texttt{DCE}}$, $\mathcal{P}_{\texttt{AMC}}$ and $\mathcal{P}_{\texttt{RAI}}$, and the maximum rounds m
for rounds $r = 1, ..., m$ **do**
  Disassemble candidate $\mathcal{C}$ into sentences $\langle s^c_1, ..., s^c_k \rangle$
  Evaluate sentence-level consistency against reference $\mathcal{R}$, and return the reasons using Eq. (1)
  Transform reasons into numeric scores using Eq. (2)
  Calculate the final consistency evaluation score $\hat{\mathcal{Z}}$ based on $\{z_1, ..., z_k\}$ using Eq. (3)
  Generate improved candidate using Eq. (4)
  Update the candidate $\langle s^c_1, ..., s^c_k \rangle \leftarrow \langle \hat{s}^c_1, ..., \hat{s}^c_k \rangle$
return $\hat{\mathcal{Z}}$, $\langle \hat{s}^c_1, ..., \hat{s}^c_k \rangle$

4 Experiments

4.1 Experimental Setup

We utilize GPT-3.5 (`gpt-3.5-turbo`) as our LLM agents, and the evaluations are carried out using the Azure OpenAI API. We employ four datasets to evaluate DCR , among which *QQP* and *PAWS* (Iyer et al., 2017) are binary datasets, whereas *SummEval* and *QAGS* utilize numeric scores to represent human preference. For a detailed description of the experimental setup please see Appendix A. For baseline methods please see Appendix B. For results on *QQP* and *PAWS* please see Table 4.

Metrics	SummEval-Consistency	
	Spearman (ρ)	Kendall-Tau (τ)
BARTScore	0.382	0.315
BERTScore	0.110	0.090
MoverScore	0.152	0.127
UniEval	0.446	0.371
GPT-Score	0.449	-
G-Eval-3.5	0.386	0.318
G-Eval-4	0.507	0.425
DCE-AMC-3.5	**0.592** (+16.76% ↑)	**0.563**(+ 32.47% ↑)

Table 1: Correlation (ρ and τ) results on SummEval.

4.2 Consistency Evaluation Results (DCE-AMC)

Summarization Consistency Evaluation. We follow the setting of previous work (Zhong et al., 2022) to evaluate different summarization consistency using summary-level Spearman (ρ) and Kendall-Tau (τ) correlation. Table 1 shows our method outperforms other baseline metrics using LLM-based evaluators. DCE-AMC-3.5, powered by GPT-3.5, even outperforms state-of-the-art methods such as G-Eval baseline using a more powerful GPT-4, by a considerable margin (+16.8% and +32.5% respectively).

Factual Consistency Evaluation. While advanced NLG models are capable of generating high-quality responses, LLMs are known to occasionally produce non-factual statements or hallucinate facts. Recent work (Manakul et al., 2023) has been conducted to identify such inconsistencies in terms of factuality. To verify the effectiveness

Metrics	QAGS-CNN			QAGS-XSUM		
	Pearson (r)↑	Spearman (ρ)↑	Kendall-Tau (τ)↑	Pearson (r)↑	Spearman (ρ)↑	Kendall-Tau (τ)↑
BERTScore	0.576	0.505	0.399	0.024	0.008	0.006
MoverScore	0.414	0.347	0.271	0.054	0.044	0.036
UniEval	0.682	0.662	0.532	0.461	0.488	0.399
G-Eval-3.5	0.477	0.516	0.410	0.211	0.406	0.343
G-Eval-4	0.631	**0.685**	0.591	0.558	0.537	0.472
DCE-AMC-3.5	**0.699**	0.648	**0.596**	**0.573**	**0.573**	**0.573**

Table 2: Pearson (r), Spearman (ρ), and Kendall-Tau (τ) correlations of different baseline metrics on QAGS-CNN and QAGS-XSUM benchmark.

Dataset (size)	SummEval (1600)		QAGS-CNN (236)		QAGS-XSUM (239)	
	Sentence	Paragraph	Sentence	Paragraph	Sentence	Paragraph
Inconsistent data	286	209	111	68	86	90
Corrected data with RAI	248	198	89	64	84	82
Consistency improvement	86.71% ↑	94.73% ↑	88.29% ↑	94.11% ↑	97.67% ↑	91.11% ↑

Table 3: Consistency improvement with RAI in one iteration across all three summarization tasks.

of our method in evaluating hallucination, we test it on the QAGS benchmark, which includes two summarization datasets: QAGS-CNN and QAGS-XSUM. Table 2 provides a comprehensive comparison of various metrics based on Pearson, Spearman, and Kendall-Tau correlations. Our proposed DCE-AMC outperforms all the baseline methods on QAGS-XSUM even with a less powerful model.

4.3 Consistency Improvement Results (RAI)

After implementing DCE and AMC, we can quantitatively determine whether each candidate is consistent (score = 1) to the reference or not (score <1). Table 3 - *Sentence* column offers a statistical analysis of the number of inconsistent data after evaluations (DCE-AMC), revealing 286, 111, and 86 inconsistent candidates for the SummEval, QAGS-CNN, and QAGS-XSUM respectively. Identifying these inconsistent candidates is valuable but the more critical objective is how to improve these responses to align with the references. To achieve this goal, we generate a new response by implementing RAI based on the reasons provided by DCE, and then use DCE to re-evaluate these improved responses. We observe a significant improvement with most inconsistencies corrected, specifically 84 out of 86 examples on the QAGS-XSUM benchmark. The rate of consistency improvement is 86.71%, 88.29%, and 97.67% on SummEval, QAGS-CNN, and QAGS-XSUM respectively. These impressive results demonstrate that our reasoning approach RAI not only provides better consistency evaluation metrics that align more closely with human judgments, but also sheds light on improving consistency beyond evaluation. This finding is particularly crucial for mitigating hallucination once we detect non-factual statements via consistency checks. It's worth noting that our reasoning method RAI is a generic component that can also be applied directly at the paragraph level, and the improvement in this context is significant as well, as illustrated in Table 3 - *Paragraph* column. Additional analysis on paragraph level are in Section D.

4.4 Analysis

Multi-round Consistency Improvement. Table 3 showcases encouraging results on consistency improvement via RAI. To bring it to another level, Fig. 3 shows multi-round consistency improvement by iteratively applying DCR . The convergence of consistency improvement is remarkably swift, achieving nearly 100% in just two rounds. The convergence rate on the QAGS datasets is highly consistent across both subsets, slightly surpassing SummEval due to its high consistency rate after the first round. This is also corroborated by the frequency distribution of the consistency score (Fig. 3 (right)). As the number of rounds increases, the lower consistency scores (<1) gradually decrease, and more inconsistent candidates tend to be consistent, where the score is 1.

Computational Cost. We assessed the computational cost of our method based on wall-clock time,

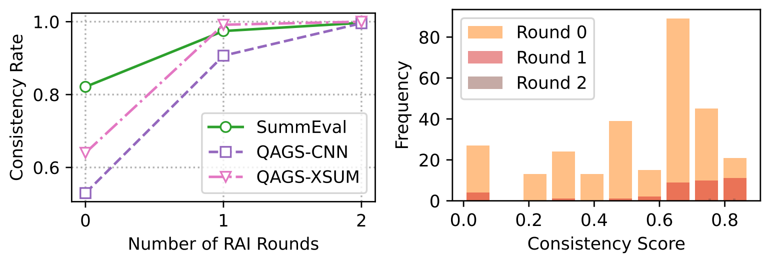

Figure 3: Multi-round consistency improvement

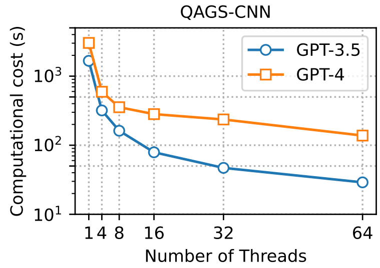

Figure 4: Computational cost.

which is primarily consumed by LLMs inference. The divide-conquer strategy we employed is highly scalable through parallelism. Fig. 4 illustrates the computational cost of GPT-3.5 and GPT-4 with varying numbers of threads on the QAGS-CNN benchmark. A clear reduction in computational cost is observed as the number of threads increases. It's important to note that the decrease in time is more significant when transitioning from a single thread to four threads, but tends to plateau as more threads are utilized. While GPT-3.5, being the smaller LLM, is a more efficient option, GPT-4 often delivers better performance.

5 Related Work

LLM-based Evaluations. Recent proposed LLM-based evaluators (Wang et al., 2023), such as GPTScore (Jinlan et al., 2023) and G-Eval (Liu et al., 2023b), have demonstrated competitive performance on multiple NLG tasks. However, these LLM evaluators often exhibit lower correlations with human judgments and may pose potential risks of producing hallucinated or overconfidence scores (Kadavath et al., 2022; Zhou et al., 2023). Our proposed DCR framework addresses these challenges through a divide-conquer strategy (DCE) coupled with a numeric score system (AMC). Our method does not rely on LLMs to directly output numeric scores, thus providing a more accurate and comprehensive score that better aligns with human feedback.

Consistency Evaluations. Consistency checking plays an essential role in a wide range of NLG tasks, including question-answering (Durmus et al., 2020; Wang et al., 2020), factual knowledge extraction (Elazar et al., 2021), summarization (Durmus et al., 2020) and hallucination detection (Manakul et al., 2023). However, due to various limitations of existing methods, such as reliance on additional pre-trained models or question sets (Durmus et al., 2020), it is highly desirable to develop a unified and automatic consistency metric (Wang et al., 2022). Our proposed framework successfully fills this gap and demonstrates superior performance compared to state-of-the-art baselines (Jinlan et al., 2023; Liu et al., 2023b; Wang et al., 2023). More importantly, our proposed RAI enables consistency improvement where the re-generated candidate response significantly helps mitigate LLM hallucinations (Dhuliawala et al., 2023; Mündler et al., 2023; Zhang et al., 2023b) in summarization, and open-book QA tasks (Li et al., 2023).

6 Industrial Application

DCR can be conveniently integrated into various industrial downstream applications, specifically in question-answering (QA), summarization, and retrieval augmented generation (RAG) tasks, in both an online and offline fashion. In an online fashion, DCR enables auto-mitigation of hallucinations by reducing inconsistency before sending responses to customers. In an offline fashion, DCR empowers consistency evaluations and hallucination detection to gauge the reliability, trustworthiness, and trends of LLM systems.

7 Conclusion

We proposed a general evaluation framework based on a divide-and-conquer strategy for assessing the consistency between the LLM-generated output and the reference texts across various NLG tasks. The proposed method can leverage analytical reasoning to generate revised text with improved consistency. Through comprehensive and systematic empirical study across multiple benchmarks in semantic, factual, and summarization consistency tasks, we demonstrated that our approach significantly outperforms existing methods in evaluating and enhancing the consistency of LLM-generated content. Despite these advancements, we acknowledge several potential limitations of our proposed method, refer to Appendix 8.

8 Limitation

Despite these advancements, we acknowledge several potential limitations of our proposed method:

Not a Silver Bullet. While our sentence-level approach (DCE-AMC) excels in evaluating *consistency* and *detecting hallucination*, it may not be universally effective for all dimensions of text evaluation, even with updated criteria in prompts. For instance, dimensions such as *coherence*, which pertains to the collective quality of all generated sentences, or *relevance*, which involves selecting important information and eliminating redundant content from the reference text, require a holistic focus on the entire candidate. These dimensions may not be ideally suited for our DCE-AMC approach. However, if a different evaluator that outputs reasons for action is used, our AMC and RAI could still be employed to quantify and improve performance on such dimensions.

Garbage in, Garbage Out. The DCR framework requires two inputs: a reference paragraph and a candidate paragraph. As we use the reference paragraph as the target for consistency and hallucination checks, any non-factual statements present in the reference paragraph would not be detected by our method. Therefore, for tasks such as retrieval-augmented generation (RAG), the accuracy of our method is inherently limited by the correctness of the input paragraphs.

Meta-prompting. Our DCR framework requires hand-craft prompts for specific tasks, and acknowledges that this is a general hurdle shared by all works relying on LLMs, which include G-Eval (Liu et al., 2023b), GPTScore (Jinlan et al., 2023), and Self-refine (Madaan et al., 2023). Specifically, in G-Eval, different prompts will need to be composed for different aspects: consistency, coherence, etc. Self-refine defines multiple customized prompts to perform their INIT - FEEDBACK – REFINE components. Our current solution is to structure our prompts in a modularized manner so task-specific content can be updated easily. However, an automated prompt-tuning procedure is beyond the focus of our study but we leave this for future work.

References

Yuan Zhang amd Jason Baldridge and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Reinald Kim Amplayo, Peter J Liu, Yao Zhao, and Shashi Narayan. 2022. Smart: Sentences as basic units for text evaluation. In *The Eleventh International Conference on Learning Representations*.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R Bowman, and Kyunghyun Cho. 2023. Two failures of self-consistency in the multi-step reasoning of llms. *arXiv preprint arXiv:2305.14279*.

Jacob Devlin, Ming-Wei Chang, Kenton aLee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints*, pages arXiv–2305.

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

Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. 2023a. Reta-llm: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

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

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. Eed: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520.

Miloš Stanojević and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the factual consistency of large language models through summarization. *arXiv preprint arXiv:2211.08412*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A Malin, and Sricharan Kumar. 2023a. Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023b. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.

A Extended Description on Experimental Setup

We utilize GPT-3.5 (`gpt-3.5-turbo`) and GPT-4 (`gpt-4`) as our LLM agents, and the evaluations are carried out using the Azure OpenAI API. We set the temperature to 0.0 to generate responses via the greedy algorithm. The specific prompts used for each LLM agent are detailed in the Appendix (from Table 9 to Table 14). All experiments are conducted on our local machine (Macbook-Pro with M1 chip) without the need for GPU resources. In our experimental setup, we set both α and β in Eq. (3) to 0. We employ four datasets to evaluate DCR where QQP and PAWS are binary datasets, as well as SummEval and QAGS have numeric scores representing human judgments.

- **QQP and PAWS**: Quora Question Pair corpus (Iyer et al., 2017) and the Paraphrase Adversaries from Word Scrambling dataset (amd Jason Baldridge and He, 2019) contain pairs of sentences labeled to indicate whether they are paraphrases or not, while PAWS specifically focuses on the adversarial paraphrases. Following the guidance of BERTScore (Zhang et al., 2020), we are using the PAWS development set and the first 5000 from the training set of QQP.
- **SummEval** (Fabbri et al., 2021) is a standard dataset that assesses various summarization evaluation techniques. It gathers human ratings in various aspects and is built on the CNN/DailyMail dataset (Hermann et al., 2015). In this study, we mainly focus on the consistency evaluation.
- **QAGS** (Wang et al., 2020) serves as a benchmark for assessing hallucinations in summarization tasks. Its objective is to evaluate the consistency aspect of summaries across two distinct summarization datasets: QGS-CNN and QAGA-XSUM.

Here we provide a detailed explanation of the "reference" used in our experiments. For Paraphrase detection tasks, such as the QQP dataset, each question pair is annotated with a binary value indicating whether the two questions are paraphrases of each other. We consider "question1" as the "reference" and "question2" as the "candidate", and our task is to evaluate if the candidate is consistent with the reference in semantic meaning. For Summarization tasks, SummEval datasets include original source articles, machine summaries, and human summaries. Our "reference" in this task is the original source article, and our "candidate" is the machine summaries. Our task is to check the factual consistency between them without relying on any additional golden reference or ground truth.

B Baseline Methods

We evaluate DCR against a variety of evaluation metrics and LLM-based evaluators that have achieved state-of-the-art performance.

- **BERTScore** (Zhang et al., 2020) calculates the similarities between two pieces of text using the contextualized embedding derived from the BERT model(Devlin et al., 2019). It operates as a similarity-based assessment tool, which has been widely used for various applications.
- **MoverScore** (Zhao et al., 2019) enhances BERTScore by incorporating soft alignments and introducing new aggregation techniques to provide a more robust similarity assessment.
- **UniEval** (Zhong et al., 2022) is a consolidated evaluator capable of assessing various elements of text generation as QA tasks. It manages diverse evaluation tasks by modifying the question format.
- **GPTScore** (Jinlan et al., 2023) is an LLM-based evaluator that assesses texts using pre-training models, e.g., GPT-3, and is designed to provide a higher likelihood to high-quality generated text.
- **G-Eval** (Liu et al., 2023b) is another LLM evaluator that utilizes LLMs with a chain-of-thoughts (CoT) approach with a form-filling paradigm to evaluate the quality of NLG outputs.

C Additional Experiments

Semantic Consistency Evaluation. Table 4 shows the Area Under the ROC curve (AUROC) for automatic baseline metrics and our method, following the practice of BERTScore (Zhang et al., 2020). We note that while most metrics from BERTScore perform acceptably on QQP, they exhibit a significant performance drop on PAWS. This suggests that these baseline metrics struggle to detect the challenging adversarial examples from a semantic consistency perspective. In contrast, our method outperforms all the baseline metrics on both QQP and PAWS, without a significant drop. Notably, `DCE-AMC` demonstrates superior robustness in adversarial paraphrase classification (semantic consistency) achieving a relatively large improvement (+1.4% in QQP and +11.1% in PAWS) compared to BERTScore.

Metrics	QQP	PAWS
BLEU (Papineni et al., 2002)	0.707	0.527
METEOR (Banerjee and Lavie, 2005)	0.755	0.532
ROUGE-L (Lin, 2004)	0.740	0.536
CHRF++ (Popović, 2015)	0.577	0.608
BEER (Stanojević and Sima'an, 2014)	0.741	0.564
EED (Stanchev et al., 2019)	0.743	0.611
CharacTER (Wang et al., 2016)	0.698	0.650
BERTScore (Zhang et al., 2020)	0.777	0.693
`DCE-AMC-3.5` (our method)	**0.788**	**0.770**

Table 4: AUROC results on QQP and PAWS

D Additional Analysis

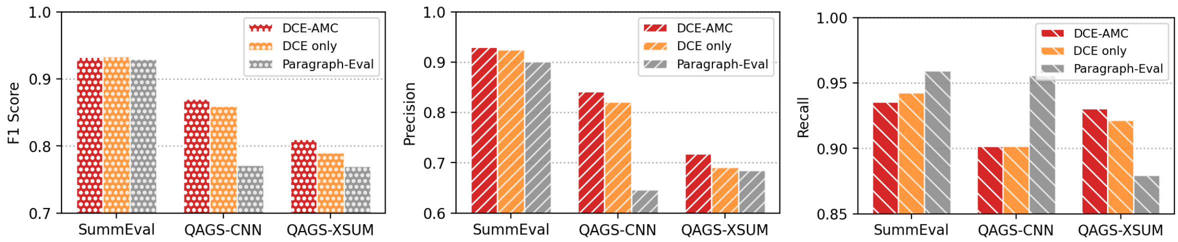

Figure 5: F1 score, precision, and recall performance of our method on sentence-paragraph and paragraph-paragraph(Paragraph Eval) evaluations.

Why `DCR` Prefers Sentence-to-Paragraph Evaluation? To further assess the potential advantage of the sentence-paragraph approach in consistency checking, we employed the same logic of outputting decisions and reasons as used in `DCE` and developed an evaluator at the paragraph-paragraph level, with prompts provided in Appendix (Table 13). The comparative results between paragraph-paragraph level and sentence-paragraph level can be viewed in Fig. 5. While the recall of paragraph-paragraph evaluation is higher on SummEval and QAGS-CNN benchmarks, its overall performance in terms of the F1 score and precision is lower than that of sentence-paragraph evaluations, particularly on the QAGS benchmark. This combination of higher recall and lower precision implies that more candidates are incorrectly marked as consistent. For consistency checking tasks, metrics with low recall and high precision (sentence-paragraph) are preferable to metrics with high recall and low precision (paragraph-paragraph), erring on the side of caution.

In addition to superior accuracy, sentence-paragraph evaluations can facilitate more thorough inconsistency remediation when integrating with `RAI`. We compared the performance improvement between our sentence-paragraph `DCE` and paragraph-paragraph, as indicated in Table 3. Despite the higher recall of the paragraph-paragraph approach, fewer items are flagged as inconsistent, resulting in fewer candidates being corrected, even though the improvement rate is higher. In fact sentence-paragraph `DCE` leads to

25.25% and 39.05% more corrections compared to the paragraph-paragraph approach in SummEval and QAGS-CNN respectively. Therefore, our sentence-paragraph approach not only outperforms in terms of F1 score and precision during consistency checks but also facilitates comprehensive improvements through RAI.

Is Auto-metric Converter Necessary? We present a comparison of our method, both with and without AMC, as shown in Fig. 5. We observe that our method with only the DCE *(red bar)* performs marginally better on the SummEval dataset but underperforms DCE-AMC *(orange bar)* on all other benchmarks. Although DCE plays a key role in our method, the AMC component is still desirable and highly necessary not only because it shows better performance, but also because it facilitates the conversion of *reasons* outputted by DCE to a numeric system. This conversion is both user-friendly and practical, making it easy for humans to understand and apply. Furthermore, it provides a straightforward means of evaluating the effectiveness of the DCE component.

RAI improvement Evaluation. To ensure the mitigated response after RAI does make sense. We randomly selected 30 revised examples and examined them manually. 2 of the cases where not all inconsistencies were migrated in one iteration, but were picked up in a second iteration. All 30 cases generate reasonable results with inconsistency reduced. Examples of the improvement can be seen in Appendix F.

The Effect of LLM models. We evaluated the DCR performance using different LLMs across all three benchmarks shown in Table 5. DCE-AMC-4 generally outperforms DCE-AMC-3.5 across all datasets. The performance gap between the two LLM models suggests that GPT-4 can further enhance performance, especially for more complex evaluation tasks. Nonetheless, the benefits of GPT-3.5, such as higher computational efficiency and lower API costs, should not be overlooked.

Metrics	SummEval		QAGS-CNN		
	Spearman (ρ)	Kendall-Tau (τ)	Pearson (r)	Spearman (ρ)	Kendall-Tau (τ)
DCE-AMC-3.5	0.592	0.563	0.699	0.648	0.596
DCE-AMC-4	**0.700**	**0.668**	**0.782**	**0.760**	**0.706**

Table 5: Effect of base LLM (GPT-3.5 vs GPT-4)

E Example of Divide-Conquer Evaluator Prompt

Your task is to evaluate whether the summary is consistent with the article. You will evaluate it by going through each sentence of the summary and check against the following procedures: • *Understands all the aspects of the sentence, and compare if each aspect exists in the article* • *If it does, compare if the information in this sentence is consistent with what is in the article* • *Compare if all the information in this sentence can be directly inferred or entailed from what is in the article. It is OK that not all information from the article exists in this summary*

Table 6: Summarization Consistency Divide-Conquer Evaluator Prompt

F Improvement Iterations Example

Original Article	... The six-time Olympic gold medallist will compete at the relay championship on May 2 and 3 as part of the Jamaican team . ' I 'm happy to be part of the Jamaican team for the IAAF / BTC World Relays in the Bahamas . I am fit , healthy and ready to run , ' said Bolt . Usain Bolt has confirmed he will be part of Jamaica 's team at the World Relays in the Bahamas Bolt reacts as he wins 4x100m gold at the London Olympic Games in 2012 ' I hear the meet was a lot of fun last year and there was a great atmosphere . Jamaica has a long and successful tradition in relays and when we put on the national colours we always do our best to make the country proud , ' he added . JAAA General Secretary Garth Gayle commented , 'We were extremely pleased that Usain was available for selection and that the world 's fastest man will be running for Jamaica . We can expect some sprint magic on the track in the Bahamas on 2nd and 3rd May . ' ...
Original Summary	Usain bolt will compete at the relay championship on may 2 and 3 as part of the jamaican team . The six-time olympic gold medalist will be part of jamaica 's team at the iaaf/btc world . Bolt is the iaaf/btc general secretary garth gayle said he always does ' best to make his country proud ' . ...
Divide-Conquer Evaluator Evaluation Reasons	• This sentence is consistent with the article. The article states that Usain Bolt will compete at the IAAF/BTC World Relays in the Bahamas next month as part of the Jamaican team • This sentence is consistent with the article. The article confirms that Usain Bolt, a six-time Olympic gold medallist, will be part of Jamaica's team at the IAAF/BTC World Relays • This sentence is not consistent with the article. The article does not state that Bolt is the IAAF/BTC General Secretary. The General Secretary is Garth Gayle. However, the article does mention that Bolt always does his best to make his country proud
First Round Improved Answer	Usain bolt will compete at the relay championship on may 2 and 3 as part of the jamaican team. The six-time olympic gold medalist will be part of jamaica 's team at the iaaf/btc world. Bolt, as stated by the IAAF/BTC General Secretary Garth Gayle, always does his best to make his country proud. ...
Divide-Conquer Evaluator Re-Evaluation Reasons	• This sentence is consistent with the article. The article states that Usain Bolt will compete at the IAAF/BTC World Relays in the Bahamas next month, which is the relay championship mentioned in the summary. The dates and his participation as part of the Jamaican team are also mentioned in the article • This sentence is consistent with the article. The article mentions that Usain Bolt, a six-time Olympic gold medallist, will be part of the Jamaican team at the IAAF/BTC World Relays. • This sentence is not consistent with the article. The article does mention that Bolt always does his best to make his country proud, but it is Bolt himself who says this, not the IAAF/BTC General Secretary Garth Gayle.'
Second Round Improved Answer	Usain bolt will compete at the relay championship on may 2 and 3 as part of the jamaican team. The six-time olympic gold medalist will be part of jamaica 's team at the iaaf/btc world. Bolt, as stated by himself, always does his best to make his country proud. ...

Table 7: SummEval Improvement Iteration Example

Original Article	Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley on Sunday . 'Just been watching the game , did you miss the coach ? RubberDub 7minutes , ' Merson put on Twitter . Merson initially angered Townsend for writing in his Sky Sports column that 'if Andros Townsend can get in (the England team) then it opens it up to anybody . ' Paul Merson had another dig at Andros Townsend after his appearance for Tottenham against Burnley Townsend was brought on in the 83rd minute for Tottenham as they drew 0-0 against Burnley Andros Townsend scores England 's equaliser in their 1-1 friendly draw with Italy in Turin on Tuesday night The former Arsenal man was proven wrong when Townsend hit a stunning equaliser for England against Italy and he duly admitted his mistake . 'It 's not as though I was watching hoping he would n't score for England , I 'm genuinely pleased for him and fair play to him ,Äì it was a great goal , ' Merson said . 'It 's just a matter of opinion , and my opinion was that he got pulled off after half an hour at Manchester United in front of Roy Hodgson , so he should n't have been in the squad . 'When I 'm wrong , I hold my hands up . I do n't have a problem with doing that - I 'll always be the first to admit when I 'm wrong . ' Townsend hit back at Merson on Twitter after scoring for England against Italy Sky Sports pundit Merson (centre) criticised Townsend 's call-up to the England squad last week Townsend hit back at Merson after netting for England in Turin on Wednesday , saying 'Not bad for a player that should be 'nowhere near the squad ' ay @ PaulMerse ? ' Any bad feeling between the pair seemed to have passed but Merson was unable to resist having another dig at Townsend after Tottenham drew at Turf Moor
Original Summary	Paul merson was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley . Andros townsend scored the tottenham midfielder in the 89th minute . Paul merson had another dig at andros townsend after his appearance . The midfielder had been brought on to the england squad last week . Click here for all the latest arsenal news .
Divide-Conquer Evaluator Evaluation Reasons	• This sentence is not consistent with the article. The article states that it was Andros Townsend, not Paul Merson, who was brought on with only seven minutes remaining in the game. • This sentence is not consistent with the article. The article does not mention Andros Townsend scoring in the 89th minute. • This sentence is consistent with the article. The article mentions that Paul Merson had another dig at Andros Townsend after his appearance. • This sentence is not consistent with the article. The article does not mention that the midfielder was brought onto the England squad last week. • This sentence is not consistent with the article. The article does not provide a link or mention to click for the latest Arsenal news.
Improved Answer	Andros Townsend was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley. Andros Townsend was brought on in the 83rd minute for Tottenham as they drew 0-0 against Burnley. Paul merson had another dig at andros townsend after his appearance . Sky Sports pundit Merson criticised Townsend's call-up to the England squad last week. The former Arsenal man was proven wrong when Townsend hit a stunning equaliser for England against Italy.

Table 8: SummEval Improvement Iteration Example 2

G Full sets of Prompts

You are an evaluator. You will be given a true answer and an attempt answer. The true answer is the ground truth answer. Attempt answer is the answer you want to evaluate. **Criteria:** Your task is to evaluate whether the attempt answer is consistent with the true answer. You will evaluate it by: * Listing all the aspects in the attempt answer * Compare if each aspect exists in the true answer * If it does, compare if the information in the attempt answer is consistent with what is in the true answer * It is OK that not all information from the true answer exists in attempt answer **Given:** ## True Answer ## {*true answer*} ## Attempt Answer ## {*answer to evaluate*} **Task** Work in a step-by-step way to make sure we get the right answer. You will format the output in JSON as follows: {"reason": [{"sentence": "original sentence", "reason": "why this sentence is or is not consistent with the true answer"}], "is_consistent": true/false} Here is the evaluation in JSON format:

Table 9: Semantic Consistency Divide-Conquer Evaluator Prompt

You are an evaluator. You will be given an article and a summary.
The summary contains a summarized version of the article.

Criteria:
Your task is to evaluate whether the summary is consistent with the article. You will evaluate it by going through each sentence of the summary and check against the following procedures:

* Understands all the aspects in the sentence, who is doing what at when and where and what are the impact etc.
* Compare if each aspect exists in the article
* If it does, compare if the information in this sentence is consistent with what is in the article
* Compare if all the information in this sentence can be directly inferred or entailed from what is in the article, including but not limited to who, what, when, where, etc.
* It is OK that not all information from the article exists in this summary

Given:
Article
{*article*}
Summary
{*summary*}

Task
Work in a step-by-step way to make sure we get the right answer. You will format the output in JSON as follows:
{"reason": [{"sentence": "original sentence", "reason": "why this sentence is or is not consistent with the article. You should start with ẗhis sentence is consistent with the articleör ẗhis sentence is not consistent with the article"}], "is_consistent": true/false}

Here is the evaluation in JSON format:

Table 10: Summarization Consistency Divide-Conquer Evaluator Prompt

You are an evaluator. You will be given a list of paragraphs about "attempt answer". Your job is to:
* Identify whether each paragraph is positive or negative
* If the paragraph is positive, mark it as 1,
* If the paragraph is negative, mark it as -1.
* Output the mark for each paragraph in a JSON array

Example
Given paragraphs:
*"The attempt answer is incorrect as it states that employees in the US are not eligible to participate in the ESPP, which contradicts the true answer. So it is incorrect",
*"The attempt answer adds a new aspect that is not in the true answer.",
*"Yet it does list the correct article. And that is helpful."

Thought:
The first paragraph is negative as it mentions the attempt answer is wrong. Thus mark -1
The second paragraph is negative as it adds something that is not in true answer. Thus mark -1
The third paragraph is positive. Thus mark 1

Answer:
{"reason": ["The first paragraph is negative as it mentions the attempt answer is wrong. Thus mark -1", "The second paragraph is negative as it adds something that is not in the true answer. Thus mark -1", "The third paragraph is positive. Thus mark +1"], "answer": [-1, -1, 1]}

Given:
Attempt Answer ##:
{*attempt answer*}

Answer:

Table 11: Auto-Metric Converter Prompt

You are a good writer. You will be given:
* An article
* A list of objects, each have two fields: sentence and reason
  ** sentence: These sentences are summaries of the given article.
  ** reason: These are the reasons why the sentence is consistent with the article or not.

Your job is to rewrite these sentences:
* If the sentence is consistent with the article, you can keep it as it is
* If the sentence is not consistent with the article, you can re-write it to make it consistent with the article based on the reasons given.

Article
{*article*}
Sentences
{*sentences*}

Task
Work in a step-by-step way to make sure we get the right answer. You will format the output in JSON as follows:
[{"sentence": "original sentence", "improved_sentence": "improved sentence", "reason": "if it is improved, how it is improved. if not, say 'ALREADY CONSISTENT'"}]

Table 12: Reason-Assisted Improver Prompt

You are an evaluator. You will be given an article and a summary.
The summary contains a summarized version of the article.

Criteria:
Your task is to evaluate whether the summary is consistent with the article. You will evaluate it by going through the summary and checking against the following procedures:

* Understands all the aspects in the summary, who is doing what at when and where and what are the impact etc.
* Compare if each aspect exists in the article
* If it does, compare if the information in this summary is consistent with what is in the article
* Compare if all the information in this summary can be directly inferred or entailed from what is in the article, including but not limited to who, what, when, where, etc.
* It is OK that not all information from the article exists in this summary

Given:
Article
{*article*}
Summary
{*summary*}

Task
Work in a step-by-step way to make sure we get the right answer. You will format the output in JSON as follows:
{"reason": {"sentence": "original sentence", "reason": "why this summary is or is not consistent with the article. You should start with T̈his summary is consistent with the articleör ẗhis summary is not consistent with the article"}, "is_consistent": true/false}

Here is the evaluation in JSON format:

Table 13: Paragraph Level Evaluator Prompt

You are a good writer. You will be given:
* an article
* a summary of the article
* a reason listing why the summary is or is not consistent with the article

Your job is to rewrite this summary:
* if the summary is consistent with the article, you can keep it as it is
* if the summary is not consistent with the article, you can rewrite it to make it consistent with the article based on the reasons given.

Article
{*article*}
Summary
{*summary*}
Reason
{*reason*}

Task
Work in a step-by-step way to make sure we get the right answer. You will format the output in JSON as follows:
{"sentence": "original summary", "improved_summary": "improved summary", "reason": "if it is improved, how it is improved. if not, say 'ALREADY CONSISTENT'"}

Table 14: Paragraph Level Reason Assisted Improver Prompt

H Divide-Conquer Evaluator Examples

True Answer	Attempt Answer	Is_Consistent	Reasons
How do I read and find my YouTube comments?	How can I see all my YouTube comments?	True	The sentence is asking the same question as the true answer, just phrased differently. Both are asking about how to view one's own comments on YouTube
What are the laws to change your status from a student visa to a green card in the US, how do they compare to the immigration laws in Canada?	What are the laws to change your status from a student visa to a green card in the US? How do they compare to the immigration laws in Japan?	False	• This sentence is consistent with the true answer as it asks about the laws to change status from a student visa to a green card in the US. • This sentence is not consistent with the true answer. The true answer asks for a comparison with the immigration laws in Canada, not Japan."
If I fire a bullet backward from an aircraft going faster than the bullet; will the bullet be going backward?	Do bullets travel faster than the speed of sound when shot from a gun? If not, is it possible? If they do, what gun and how much devastation occurs?	False	• This sentence is not consistent with the true answer because it is asking about the speed of bullets in relation to the speed of sound, not about the direction of a bullet when fired from a fast-moving aircraft. • This sentence is not consistent with the true answer because it is asking about the possibility of bullets traveling faster than the speed of sound, not about the direction of a bullet when fired from a fast-moving aircraft. • This sentence is not consistent with the true answer because it is asking about the type of gun and the amount of devastation, not about the direction of a bullet when fired from a fast-moving aircraft.

Table 15: QQP Evaluation Example

True Answer	Attempt Answer	Is_Consistent	Reasons
What were the major effects of the cambodia earthquake , and how do these effects compare to the Iquique earthquake in 1877 ?	What were the major effects of the Iquique earthquake , and how do these effects compare to the cambodia earthquake in 1877 ?	False	The sentence is not consistent with the true answer because it reverses the order of the earthquakes. The true answer asks about the effects of the Cambodia earthquake and compares it to the Iquique earthquake in 1877, while the attempt answer asks about the effects of the Iquique earthquake and compares it to the Cambodia earthquake in 1877.
How do I start learning electronic music production ? Should I begin with books or videos ? I have logic pro 9 and I already made some songs but I would like to start from scratch (as if I did not know anything) .	How do I begin learning electronic music production ? Should I start with books or videos ? I have logic pro 9 and I already made some songs but I would like to start from scratch (as if I did not know anything) .	True	• This sentence is consistent with the true answer. The only difference is the use of the word 'begin' instead of 'start', but they mean the same thing. • This sentence is consistent with the true answer. It is exactly the same. • This sentence is consistent with the true answer. It is exactly the same.

Table 16: PAWS Evaluation Example

Original Article	Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley on Sunday . 'Just been watching the game , did you miss the coach ? # RubberDub # 7minutes , ' Merson put on Twitter . Merson initially angered Townsend for writing in his Sky Sports column that 'if Andros Townsend can get in (the England team) then it opens it up to anybody . ' Paul Merson had another dig at Andros Townsend after his appearance for Tottenham against Burnley Townsend was brought on in the 83rd minute for Tottenham as they drew 0-0 against Burnley Andros Townsend scores England 's equaliser in their 1-1 friendly draw with Italy in Turin on Tuesday night The former Arsenal man was proven wrong when Townsend hit a stunning equaliser for England against Italy and he duly admitted his mistake . 'It 's not as though I was watching hoping he would n't score for England , I 'm genuinely pleased for him and fair play to him ,Äì it was a great goal , ' Merson said . 'It 's just a matter of opinion , and my opinion was that he got pulled off after half an hour at Manchester United in front of Roy Hodgson , so he should n't have been in the squad . 'When I 'm wrong , I hold my hands up . I do n't have a problem with doing that - I 'll always be the first to admit when I 'm wrong . ' Townsend hit back at Merson on Twitter after scoring for England against Italy Sky Sports pundit Merson (centre) criticised Townsend 's call-up to the England squad last week Townsend hit back at Merson after netting for England in Turin on Wednesday , saying 'Not bad for a player that should be 'nowhere near the squad ' ay @ PaulMerse ? ' Any bad feeling between the pair seemed to have passed but Merson was unable to resist having another dig at Townsend after Tottenham drew at Turf Moor .
Original Summary Divided into Bullet Points	• Paul merson was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley. • Andros townsend scored the tottenham midfielder in the 89th minute . • Paul merson had another dig at andros townsend after his appearance . • The midfielder had been brought on to the england squad last week . • Click here for all the latest arsenal news news .
Reasons	• This sentence is not consistent with the article because it was Andros Townsend, not Paul Merson, who was brought on with only seven minutes remaining in the game. • This sentence is not consistent with the article because it does not mention that Andros Townsend scored in the 89th minute. Also, the phrase 'scored the Tottenham midfielder' is not clear. • This sentence is consistent with the article. The article mentions that Paul Merson had another dig at Andros Townsend after his appearance. • This sentence is not consistent with the article because it does not mention that Andros Townsend was brought onto the England squad last week. • This sentence is not consistent with the article because the article does not provide any latest Arsenal news.
Is_Consistent	False

Table 17: SummEval Evaluation Example

Original Article	London 's first history day will be held on the anniversary of big ben 's first day in operation . It will be first celebrated on 31 may in 2017 with celebrations and events run by historic england . The date was decided upon after a poll involving 1,000 londoners . It was closely followed by 5 september - the date of the great fire of london . The yougov questionnaire also declared the houses of parliament as the building that best sums up london . People voted for the queen as their favourite historic london hero for the moment she secretly joined the crowds to celebrate victory in europe day . The results of the poll were released to mark the launch of historic england 's " keep it london " campaign . People were asked to select a date to celebrate the capital 's history , their historic hero and the building that sums up london . Big ben 's first day in operation was 31 may 1859 . The campaign is intended to encourage londoners to notice , celebrate and speak up for the heritage of their city , historic england said . The public body has also launched a film entitled i am london , which celebrates the historic buildings and places that have borne witness to the capital 's history . Duncan wilson , chief executive of historic england , said : " our heritage is a source of pride , a reminder of the city 's past , a foundation for its present and the building blocks for its future .
Original Summary Divided into Bullet Points	• Big ben 's 150th anniversary has been chosen as the date to celebrate london 's history .
Reasons	• This sentence is not consistent with the article. The article states that London's first history day will be held on the anniversary of Big Ben's first day in operation, which is 31 May 1859. However, it does not mention that this date is the 150th anniversary of Big Ben.
Is_Consistent	False

Table 18: QAGS-XSUM Evaluation Example

Original Article	A southern iowa chiropractor accused of accepting sex as payment for his services and performing exorcisms on patients has surrendered his state license . The iowa board of chiropractic released a report wednesday detailing charges against charles manuel , of lamoni . Manuel signed an agreement last month admitting his misdeeds and pledging not to apply for reinstatement for at least 10 years . Patient satisfaction : a chiropractor in iowa has surrendered his license to practice and admitted to swapping services for sex and performing exorcisms on some patients . Nonetheless , he 's received outstanding evaluations for patient satisfaction on healthgrades . Com . The agreement requires that manuel prove any circumstances surrounding the revocation of his license no longer exist before he can resume practicing chiropractic in the state . Those circumstances included bartering sex for services with some patients . Manuel also recommended that patients stop taking medication he prescribed to them . A woman who answered a call to manuel 's home from the des moines register declined to comment on the case . A woman at his former practice said he had n't worked there for some time . A lamoni address listed on the yelp page for manuel 's practice appears to be a home on a residential street . While maneul has received just three patient survey responses on healthgrades . Com , those responses were quite positive . The disgraced chiropractor received a perfect five out of five stars in patient satisfaction . Strange practice : charles manuel , who admitted wrongdoing to the iowa board of chiropractic , listed his practice 's addresses on this residential street in the small agricultural town of lamoni in southern iowa .
Original Summary Divided into Bullet Points	• A chiropractor in iowa has surrendered his license to practice and admitted to swapping services for sex and performing exorcisms on some patients. • Manuel also recommended that patients stop taking medication no longer exist before he can resume practicing chiropractic in the state . • The disgraced chiropractor received a perfect five out of five stars in patient satisfaction .
Reasons	• This sentence is consistent with the article. The article states that Charles Manuel, a chiropractor in Iowa, surrendered his license and admitted to accepting sex as payment for his services and performing exorcisms on patients. • This sentence is not consistent with the article. The article does mention that Manuel recommended patients stop taking medication he prescribed to them, but it does not state that this condition must be met before he can resume practicing. • This sentence is consistent with the article. The article mentions that despite his misdeeds, Manuel received outstanding evaluations for patient satisfaction on healthgrades.com, including a perfect five out of five stars.']
Is_Consistent	False

Table 19: QAGS-CNN Evaluation Example

I Example of setting α and β in equation 3

α and β are added to ensure we are using a fully sentence level analysis. Below we give an example on how to use them to achieve this goal.

Suppose the output from Divide-Conquer Evaluator with customized prompt is:

```
{
  "is_consistent" : False,
  "reasons" : [
    "The two paragraphs are not consistent.",
    "This sentence is consistent."
    "This sentence is not consistent."
  ]
}
```

After inspecting the *reasons* we notice that the first entry is not a sentence level analysis, but a paragraph level one. Thus when calculating scores we should remove the impact of this line.

Since in this case the first entry is negative, it will be given -1 score, thus we should set $\alpha = 1$ to mitigate this effect. Similarly, because we only have one entry that is not sentence level, we would set $\beta = -1$.

Thus for this particular case, equation 3 will be:

$$\mathcal{Z} = \left(\sum_{i=1}^{k} z_i + 1 \right) / (k-1), \quad \hat{\mathcal{Z}} = (\mathcal{Z} + 1)/2, \quad \hat{\mathcal{Z}} \in [0, 1] \tag{5}$$

J What does LLM think of sentence level approach?

To have some more fun, we asked different LLMs on what they think the best approach to evaluate consistency is, token level, paragraph level, or sentence by sentence level. We have a unanimous decision :)

Input Prompt	*Given a reference, we have a generated paragraph from LLM. Our objective is to check whether the generated text is consistent to the reference or not. Can you you tell me, to achieve the best accuracy, would you choose to compare them token-by-token, direct check of paragraphs as a whole, or sentence by sentence check on the semantic meaning?*
GPT 3.5	To achieve the best accuracy in checking the consistency of generated text with a reference, it is recommended to compare them sentence by sentence on the semantic meaning. This approach allows for a more nuanced analysis of the generated text and can identify inconsistencies that may not be apparent when comparing them token-by-token or as a whole paragraph. Additionally, comparing them sentence by sentence allows for a more efficient and targeted review of the generated text, as inconsistencies can be identified and addressed on a sentence-by-sentence basis.
GPT 4	To achieve the best accuracy, it would be most effective to check sentence by sentence on the semantic meaning. This is because language can be complex and nuanced, and a direct token-by-token or whole paragraph comparison might miss subtle differences in meaning. By checking sentence by sentence, you can ensure that the generated text accurately reflects the reference in both content and context.
PaLM2	To achieve the best accuracy in checking whether a generated paragraph from LLM is consistent with a reference, it would be best to compare them sentence by sentence on the semantic meaning. This is because comparing token-by-token would be too time-consuming and would not take into account the overall meaning of the sentences. Direct check of paragraphs as a whole would also not be as effective, as it would not identify specific inconsistencies between the sentences. Comparing them sentence by sentence on the semantic meaning would allow for a more thorough and accurate check of the consistency between the generated paragraph and the reference.
Claude 2	Overall, I would recommend the sentence-by-sentence semantic comparison approach. By checking the semantic consistency at a sentence level, it provides a good balance between catching inconsistencies in meaning while allowing for minor paraphrasing and word order changes. The paragraph-level comparison is too holistic while the token-by-token approach is too strict. Focusing on sentence-level meaning comparisons allows catching semantic inconsistencies with reasonably high accuracy.

Table 20: LLM's opinion on our sentence level approach