Human-Aligned Long-Form Evaluation (HALF-Eval): Framework for Assessing AI-Generated Content and Improvement

Sulbha Jain
suljain@amazon.com
Amazon
Seattle, WA, USA

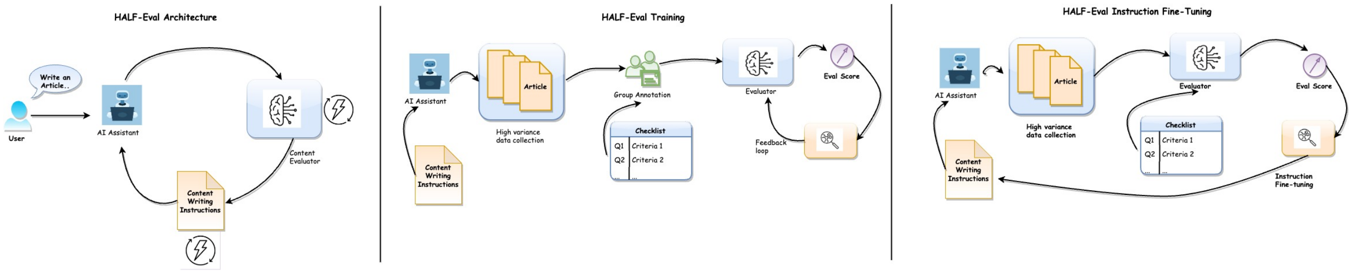

Figure 1: HALF-EVAL: Architecture diagram

Abstract

Evaluating long-form AI-generated content remains challenging due to the lack of standardized methodologies that robustly align with human judgment across formats such as articles, blogs, and essays. We introduce HALF-Eval, a scalable framework that combines structured, checklist-based evaluation with machine learning aggregation to assess key quality dimensions, including creativity, impact, coherence and relevance. Our approach leverages regression models trained on human-annotated data to synthesize checklist scores into holistic quality classifications, enabling automated yet human-aligned assessments. Experimental results demonstrate that HALF-Eval improves the quality of generated articles by 16% and blogs by 13%, while generalizing effectively to essays. The framework delivers actionable feedback for content refinement and maintains interpretability through its checklist structure. HALF-Eval advances human-centric evaluation systems and offers a robust foundation for scalable quality control in AI-generated long-form content.

CCS Concepts

• **LLM Evaluation**; • **Computing methodologies** → *Natural language processing*;

Keywords

AI-generated content evaluation, Human-aligned assessment, Long-form content, Scalable evaluation framework, Checklist-based evaluation

ACM Reference Format:
Sulbha Jain. 2025. Human-Aligned Long-Form Evaluation (HALF-Eval): Framework for Assessing AI-Generated Content and Improvement. In *Proceedings of LLM4ECommerce Workshop at the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (LLM4ECommerce Workshop at KDD '25).* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large Language Models (LLMs) have achieved significant success across diverse natural language processing (NLP) tasks, including natural language generation. Recent advancements have spurred interest in generating long-form text, which encompasses written, audio, or visual materials that exceed conventional length standards [Bai et al. 2024b; Zhou et al. 2023]. Written long-form content, for example, typically surpasses 1,000 words and is valued for its depth and comprehensive analysis [Yang et al. 2023]. Strategically incorporating long-form content into digital communication yields several benefits. From a search engine optimization (SEO) perspective, it enhances visibility by increasing audience engagement and improving search rankings [Ouyang et al. 2022]. Additionally, it builds trust with audiences by establishing the content creator as a thought leader through high-quality, authoritative information [Pavlik 2023]. The versatility of long-form content allows for repurposing across multiple platforms, such as infographics or video snippets, and contributes to a valuable resource library that aids both current and potential customers in making informed decisions [Pavlik 2023]. Its depth and lasting relevance ensure sustained engagement and traffic over time.

AI-driven tools have revolutionized long-form content creation by enabling efficient generation of blog posts, articles, and other

LLM4ECommerce Workshop at KDD '25, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

formats. These tools, based on transformer architectures [Vaswani et al. 2023], produce lengthy content at unprecedented speeds, making them ideal for organizations seeking scalable solutions to meet growing content demands [An et al. 2023]. However, while AI excels in speed and scalability, it often lacks the nuanced understanding required for high-quality content creation. Despite these limitations, AI offers advantages such as reduced operational costs and optimized content performance through integrated SEO strategies [Yang et al. 2023].

To advance research in long-context modeling, establishing reliable benchmarks for evaluating LLM performance in generating long-form text is critical. Such benchmarks facilitate systematic assessment and comparison of different models' capabilities in handling extended content. This study introduces **HALF-Eval** (Human-Aligned Long-Form Evaluation), a framework designed to assess creative long-form content, including articles and blogs. HALF-Eval establishes an open-ended methodology inspired by Bloom's Taxonomy to evaluate texts exceeding 1,000 words [Conklin 2005]. Its contributions include checklist-based evaluations aligned with human assessment criteria [Banerjee and Lavie 2005] and scalable methodologies for new content types. The framework is readily adaptable to non-English languages, thereby contributing to the broader field of natural language processing.

The key contributions of this work are as follows:

- A robust and deterministic checklist-based evaluation method for assessing open-ended content.
- Alignment of evaluation criteria with human judgment standards to improve assessment fidelity.
- A data-driven feedback mechanism designed to iteratively enhance AI-generated content.
- Scalable and generalizable evaluation methodologies capable of supporting emerging content types.

The remainder of this paper is organized as follows: Section 2 reviews the significance of long-form content in digital communication and existing evaluation methods. Section 3 outlines the human-aligned strategy and experimental design for assessing long-form content, along with evaluation metrics and performance results. Section 5 summarizes the findings and discusses future research directions.

2 Related Work

The evaluation of long-form AI-generated content builds on methodologies spanning content creation frameworks, benchmarking standards, and quality assessment techniques. Prior work in structured content generation emphasizes three phases: hypothesis formulation, topic planning, and iterative production [An et al. 2023], with storytelling identified as critical for enhancing engagement [Shao et al. 2024]. Post-production refinement requires rigorous editing for accuracy [Yang et al. 2023] and cultural localization through expert collaboration [Ouyang et al. 2022], underscoring the complexity of aligning automated systems with human standards.

Recent benchmarks such as SCROLLS [Shaham et al. 2022], ZeroSCROLLS [Shaham et al. 2023], and LongBench [Bai et al. 2024a] have advanced the evaluation of long-context LLM capabilities, yet face challenges like data contamination [Golchin and Surdeanu 2024]. Complementary frameworks like L-Eval [An et al. 2023] and BigGenBench [Kim et al. 2024b] extend evaluation to diverse domains but struggle with subjective quality dimensions. While analogy-based methods [Wijesiriwardene et al. 2023] and chunked evaluation [Ivgi et al. 2022] address specific aspects, they often neglect holistic content coherence.

Factuality assessment has seen progress through multi-step verification pipelines [Kim et al. 2024a; Wei et al. 2024] and atomic evaluation metrics [Min et al. 2023], though these methods are resource-intensive and poorly suited for creative content. Checklist-driven approaches [Que et al. 2024; Tan et al. 2024] offer practical solutions but lack scalability across content types. The continued use of metrics such as BLEU [Papineni et al. 2002], despite known limitations in aligning with human judgment [Wang et al. 2023], underscores the importance of developing human-grounded evaluation frameworks.

HALF-Eval addresses these gaps by integrating checklist-based quality dimensions (depth, coherence, relevance) with human-aligned ML aggregation. Unlike prior work focused on factuality [Rosati et al. 2024] or task-specific benchmarks [Dong et al. 2024], our framework provides a scalable, extensible methodology for diverse long-form content while mitigating biases inherent in LLM-dependent evaluation [Xu et al. 2023].

3 Experimental Setup

Our experimental setup is designed to rigorously evaluate the HALF-Eval framework for long-form AI-generated content. The HALF-Eval framework employs a systematic approach to evaluate long-form AI-generated content through a two-stage evaluation process. Our experimental methodology is designed to ensure robust measurement and improvement of content quality, even with limited initial data.

Content Generation and Evaluation Stages: Stage 1 involves comprehensive content assessment through a six-dimension checklist covering creativity, interest, coherence and relevance qualities. This evaluation is conducted either by trained human annotators or LLM judges. Stage 2 utilizes a lightweight regressor that learns weights over these facet scores to produce both a numerical quality rating (1-10 scale) and a binary high/low classification.

The process is structured into the following phases:

1. **Content Generation**: We generate long-form texts using state-of-the-art LLMs and standardized prompts spanning multiple genres.
2. **Human Evaluation**: Trained annotators independently assess each sample using a structured checklist framework.
3. **Metrics**: HALF-Eval evaluates content across four key dimensions—Creativity, Interest, Coherence, and Relevance—as well as an aggregated quality score.
4. **Model Training and Validation**: Regression models are trained on annotated data and validated on held-out samples to predict overall content quality.
5. **Prompt Fine-Tuning and Content Improvement**: Evaluation feedback is used to iteratively refine content generation prompts.
6. **Scaling to Emerging content types**: Extend evaluation framework across diverse content types.

The framework's design prioritizes reproducibility and cost efficiency by combining inexpensive LLM scoring with a simplified linear model architecture. This approach enables effective quality assessment while maintaining practical implementation requirements for real-world applications. This methodology provides a structured foundation for measuring and improving AI-generated content quality, offering a balanced approach between comprehensive evaluation and practical implementation constraints.

3.1 Content Generation

We constructed a multi-format corpus through systematic sampling and augmentation. High-quality base content was generated using Claude-Sonnet 3.5 (temperature=1.0, topP=0.9) [cla 2024], resulting in 200 articles and blogs (1,000–2,000 words) from diverse prompts. Negative samples were produced via three degradation strategies: (1) semantic perturbations through contradictory LLM rewrites (see Appendix .5), (2) automated linguistic modifications—including 15% character-level spelling errors, case randomization, and paragraph shuffling—and (3) collection of open-access web texts with inherent quality variance. The dataset was further augmented with 200 essays from the ASAP Automated Essay Scoring dataset[1], normalized to 1–10 ratings using min-max scaling.

The final balanced corpus equally represents four categories: high-quality LLM-generated content, LLM-perturbed degraded outputs, linguistically modified texts, and web-sourced materials. All texts were converted to 512-dimensional embeddings using `all-mpnet-base-v1` sentence transformers [Reimers and Gurevych 2019], and checklist scores were standardized via RobustScaler. Stratified splitting preserved category distributions across partitions: 70% for training, 20% for testing, and 10% for validation. Human quality annotations (see Section 3.2) provided ordinal supervision labels (1–10 scale) aligned with subjective quality judgments.

3.2 Checklist-Based Human Annotation

To reduce subjectivity and enhance reproducibility, we follow a structured checklist for each evaluation criterion. Annotators assign scores on a standardized scale, guided by detailed rubrics to ensure consistent application across diverse content types. This approach enables granular feedback and supports both human and automated aggregation.

A team of trained annotators evaluates a representative sample of AI-generated articles, blogs, and essays. Their checklist scores serve as ground-truth labels for training regression models, which aggregate individual criterion scores into holistic quality classifications. This machine learning component enables scalable, automated evaluation while maintaining alignment with human judgment.

Each sample is independently evaluated by three English-proficient annotators, compensated at $50 per hour. Annotators provide both continuous overall quality scores and detailed per-criterion assessments (e.g., coherence, relevance, evidence support) using standardized rubrics. Inter-annotator agreement is assessed using Cohen's κ [Cohen 1960] and intra-class correlation coefficients (ICCs) [Shrout and Fleiss 1979] (see Appendix .4). Final scores are computed as the arithmetic mean of annotator ratings, with missing ratings imputed conservatively as 1.0. While some scoring disagreements remain (a direction for future work), the processed labels exhibit sufficient consistency for effective model training.

3.3 Metrics and Analysis

HALF-Eval structures evaluation using a Key Performance Indicator (KPI) framework, categorizing criteria based on established heuristic text generation practices [Venkatraman et al. 2025]. Each content sample is systematically evaluated on:

- **Creativity and Originality**: Uniqueness of ideas and creative elements.
- **Interest and Impact**: Ability to capture and sustain reader attention.
- **Coherence and Consistency**: Logical flow, narrative structure, and internal consistency.
- **Relevance**: Alignment with the prompt or intended topic.
- **Overall**: Aggregated quality of the presented information.

These KPIs are evaluated with flexible checklist questions for each content type as listed in `Appendix Section .2`. We also report improvements in content quality via aggregated checklist scores and inter-annotator agreement. Comparative analyses across content types demonstrate generalizability.

For model performance, we use **Mean Absolute Error (MAE)** when ground-truth annotations are available:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i^{\text{pred}} - y_i^{\text{test}}| \tag{1}$$

where N is the number of samples, y_i^{pred} the predicted score, and y_i^{test} the human-annotated score. Lower MAE indicates closer alignment with human judgment.

When ground-truth labels are unavailable, we report the **Positive-Content-Rate (PCR)**, the proportion of content classified as high quality:

$$PCR = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(label_i = 1) \tag{2}$$

where $\mathbb{I}$ is the indicator function and $label_i$ the predicted class. We also generate granular KPI-related metrics (e.g., originality, coherence, relevance) to support robust empirical analysis.

3.4 Training Pipeline

Our training pipeline predicts long-form content quality by leveraging human-annotated checklist scores and overall assessments as supervision signals. Each sample is encoded using structured checklist features and semantic embeddings from both content and prompts. The primary regression model learns the mapping between checklist criteria (c_i) and overall human quality scores:

$$\text{content\_score}(S) = \sum_{i=1}^{n} w_i c_i + \epsilon \tag{3}$$

where w_i are learned weights for each checklist item (scaled 1–10), and ϵ is the residual error.

We systematically evaluate four feature configurations: (i) checklist scores only, (ii) content text embeddings with checklist scores, (iii) prompt embeddings with checklist scores, and (iv) content text embeddings alone. Multiple model architectures are benchmarked,

[1] https://www.kaggle.com/competitions/asap-aes/data

including Linear Regression, Random Forests, Gradient Boosting, and Feed-Forward Neural Networks, all initialized with default hyperparameters.

Feature importance analysis identifies which checklist criteria most strongly influence predicted quality, informing future feedback loops. For classification, quality labels are binarized using the median overall score from the training set. The resulting models support both continuous score prediction and binary classification, enabling scalable, automated evaluation of generative AI outputs.

3.5 Scoring Pipeline

To enable automated quality assessment when human-annotated checklist scores are unavailable, we introduce a two-step scoring framework based on the LLM-as-a-Judge (LLMJ) paradigm [Zheng et al. 2023]. The pipeline first uses LLMJ (Nova-Pro, Temperature=0.05) [Services 2025] to predict checklist scores for each content sample via a comprehension-style prompt (see Appendix .6). Predicted checklist scores are then input into the pre-trained regression model, which aggregates them to produce an overall quality score and binary high/low quality label.

LLMJ outputs are structured for compatibility and interpretability. While this approach enables scalable evaluation, it introduces challenges such as API rate limits and output variability. We validate the hybrid pipeline by comparing LLMJ-derived scores with human-annotated ground truth, quantifying alignment using established metrics [Roucher 2025].

3.6 Prompt Fine-tuning

After training a robust evaluator, we iteratively improve content quality by refining prompts. As shown in Figure 1 (last panel), content batches are generated and evaluated, and writing instructions are updated to address gaps in low-performing dimensions (e.g., if "Originality" scores are low, prompts are adjusted to emphasize originality).

3.7 Evaluation Generalization

To extend our evaluation framework across diverse content types including reports, newsletters, and essays we introduce three scalable methodologies. These approaches leverage both machine learning and LLM techniques to address variability in style, tone, and structure. Initial models are developed for high-priority content types, with scaling strategies as follows: (i) *Nearest Neighbor Mapping*, assigning new types to the most similar evaluated category [Lau and Biedermann 2020]; (ii) *Similarity Ranking*, generating weighted aggregate scores from the top-k most similar types; and (iii) *Content Clustering*, grouping content into clusters for robust generalization.

This design ensures the HALF-Eval framework maintains human-aligned quality standards as it adapts to new content types, with periodic updates to reflect evolving LLM capabilities and content trends.

All feature pre-processing and supervised model training were performed on CPU resources. The scoring pipeline leverages Amazon Bedrock for LLM inference and is executed on an AWS SageMaker Studio notebook instance (`ml.t3.large`, CPU).

4 Results

Section 4.1 presents evaluation results for Articles, while Section 4.2 and Section 4.3 report findings for Blogs and Essays, respectively, focusing on the performance of scalable evaluation methodologies.

4.1 Article

For non-fiction educational and informative articles, model performance was evaluated using checklist scores (Q1–Q7) as primary features. As shown in Table 4, models trained exclusively on these checklist scores achieved a mean absolute error (MAE) of 0.82, outperforming models that relied on text or prompt embeddings alone. Among the regressors tested, the Random Forest model yielded the lowest MAE (see Table 5), effectively capturing non-linear feature interactions and demonstrating efficient utilization of the checklist-based inputs.

An analysis of feature importance (Figure 8) identified Q6 "sufficient length, detail, and evidence" as the most significant predictor of article quality. This finding underscores the critical role of comprehensiveness and evidence-based reasoning in high-quality educational content. The predicted checklist scores were subsequently aggregated by the trained regression model to produce overall quality scores. Applying a threshold at the 50th percentile (score = 7.0), articles were classified as either high or low quality.

As illustrated in Figure 2, the score distributions generated by HALF-Eval closely mirror the broader variance observed in human annotations, in contrast to the narrower and generally more lenient ratings produced by LLMJ. This broader distribution enables more effective discrimination of lower-quality content, a distinction less apparent in LLMJ evaluations. However, as summarized in Table 9, MAE analysis indicates that HALF-Eval does not further reduce prediction error compared to the baseline. Notably, checklist-based feedback primarily highlighted deficiencies in Q6 (insufficient detail) and Q4 (weak coherence), providing actionable insights for iterative content refinement.

Prompt Fine-tuning: As indicated in Table 1, content quality was particularly low in the areas of interest and coherence. To address these gaps, the content generation prompt was refined, resulting in a 16% improvement in overall quality. For example, the revised prompt explicitly emphasized the need for sufficient length, thorough detail, and comprehensive evidence:

> “Is the generated content not only sufficiently long and complete but also thoroughly detailed, ensuring each argument is extensively explained and supported by comprehensive evidence?”

4.2 Blog

Blogs, which are typically characterized by an informal and conversational tone, were evaluated using the checklist-based methodology outlined in Section 3.3. Among the regression models tested, Linear Regression achieved the highest predictive accuracy for blog quality, with a mean absolute error (MAE) of 0.075 (see Table 7). Feature importance analysis (Figure 9) revealed that Q2—"depth vs. generic content"—was the most influential factor in determining perceived blog quality. This finding underscores the value of substantive, non-generic perspectives in effective blog writing. To

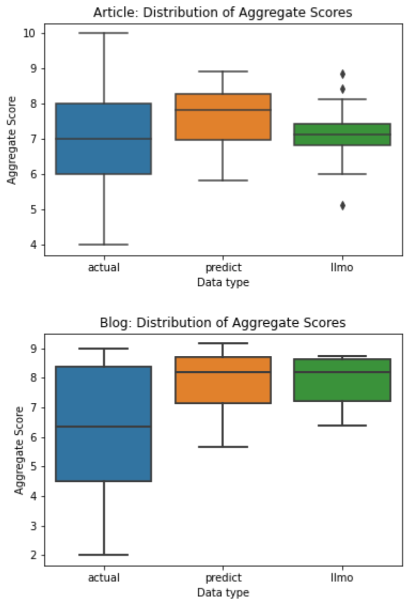

Figure 2: Score distributions for Articles [left] and Blogs [right] comparing Human Annotations, HALF-Eval, and LLMJ.

Table 1: HALF-EVAL: Content quality measurement: (B)efore and (A)fter prompt Fine-tuning for Article, Blog and Essay.

	Article		Blog		Essay	
	B	A	B	A	B	A
Creativity and Originality	8.08	9.0	7.38	7.88	4.77	8.85
Interest and Impact	6.76	8.36	8.47	9.03	3.88	8.22
Coherence and Consistency	6.67	8.89	7.44	8.5	3.84	8.26
Relevance	7.81	9.32	8.25	9.16	4.65	8.98
Overall	7.56	8.78	7.83	8.98	4.70	8.49
PCR (%)	75.0	100.0	62	100	0	100

distinguish between high- and low-quality content, a classification threshold of 7.33 was established based on the distribution of human-aligned scores.

Each blog post’s checklist scores were processed through the trained regression model to generate overall quality predictions. As shown in Figure 2[right], HALF-Eval’s predicted score distribution closely matched that of human annotations and outperformed LLMJ, which tended to assign inflated ratings (see Table 1). The integration of human-driven aggregation and calibration methods resulted in a lower MAE compared to naive LLMJ averaging, highlighting the importance of calibration for reliable automated evaluation. Qualitative analysis of prediction errors indicated that low-quality blogs frequently suffered from generic perspectives and a lack of novel insights (Q2), providing actionable feedback for content refinement.

Prompt Fine-tuning: Initial evaluations identified creativity and coherence as areas in need of improvement. To address these issues, the prompts were refined to explicitly encourage the inclusion of unique perspectives and concise, naturally structured sentences. These prompt enhancements led to a substantial increase in content quality, with average scores rising from 7.83 to 8.93—a 13% improvement.

4.3 Scaling Results for Essays

Essays, characterized by their structured argumentation and evidence-based analysis, present unique evaluation challenges compared to blogs or articles, primarily due to their emphasis on logical rigor and analytical depth. To address these challenges, we validated our scalable evaluation framework on essay content using three distinct methodologies:

1. *Nearest Neighbor Mapping:* Essays were classified as most analogous to articles, which resulted in a marginally higher MAE compared to evaluations performed with article-specific LLMJ models.
2. *Similarity Ranking:* Accuracy was improved by weighting predictions, assigning 70% influence to article models and 30% to blog models, thereby leveraging the strengths of both content types.
3. *Content Clustering:* This approach reinforced structural parallels between essays and articles by grouping them within the same content clusters, facilitating more robust evaluation.

As shown in Table 10, the Similarity Ranking-based HALF-Eval scaling approach achieved a lower MAE than LLMJ-based evaluations. These results demonstrate that similarity-driven methodologies effectively capture essay-specific qualities such as argumentative depth and evidentiary rigor. This aligns with recent research emphasizing the necessity for evaluation frameworks that accommodate the unique structural and analytical demands of academic writing.

Prompt Fine-tuning: Initial assessments indicated low content quality across all key performance indicators (KPIs). To address this, the HALF-Eval feedback loop was utilized to refine prompt instructions, with a focus on enhancing interest and coherence. The revised prompts included questions such as: Is there sufficient depth, or is the content too generic?”, Does the writing maintain a clear and engaging tone?”, and “Is the generated content consistently engaging, highly original, novel, and compelling to readers?” Implementation of these prompt updates led to a substantial improvement in content quality, with average scores increasing from 4.7 to 8.49.

4.4 Other Validation

Statistical Validation: To assess the alignment between model predictions and human annotations, we conducted a Wilcoxon signed-rank test comparing the distributions of predicted and human-assigned scores. For both articles and essays, no statistically significant differences were observed ($p > 0.05$; see Table 9) in overall score from HALF-EVAL, supporting the null hypothesis (H_0) of distributional equivalence. In contrast, for blogs, the null hypothesis was rejected ($p < 0.05$), suggesting that the model’s performance is limited by insufficient training data, particularly in the lower score regions.

Ungrounded Evaluation: HALF-Eval feedback identified recurrent deficiencies in specific key performance indicators (KPIs), such as "interest" and "coherence" for articles, "creativity" and "coherence" for blogs, and nearly all KPIs for essays. These insights informed targeted revisions to the content generation instructions, with an emphasis on increasing length, detail, and overall substantive quality. Following these prompt refinements, the Positive Content Rate (PCR) improved substantially across all content types: increasing from 75% to 100% for articles, from 62% to 100% for blogs, and reaching 100% for essays.

In summary, the framework's strong alignment with human judgment across content types establishes a scalable foundation for evaluating diverse AI-generated formats, addressing the ongoing need for robust, domain-sensitive assessment in generative AI research.

5 Conclusion

This work presents a checklist-driven evaluation framework that achieves robust alignment with human assessment standards, enabling precise and scalable classification of AI-generated long-form content. The Human-Aligned Long-Form Evaluation (HALF-Eval) system advances automated content assessment through two key contributions: (i) human-calibrated weighting mechanisms that prioritize evaluation criteria based on empirical importance, (ii) the integration of human judgment patterns via prompt fine-tuning for nuanced quality improvement. Notably, human-annotated checklist scores are required only during the training phase, allowing the framework to operate fully automatically when scoring new, unseen content. Empirical results demonstrate substantial improvements in content quality 13% for articles and 16% for blogs while maintaining generalizability across diverse content types, such as essays.

These findings have important implications for the development of future AI content generation and evaluation systems. The success of checklist-based, human-aligned assessment in capturing nuanced quality dimensions including depth, coherence, and relevance provides a reproducible blueprint for evaluating emerging content formats. Furthermore, the demonstrated effectiveness of similarity-ranking methods for essay evaluation underscores the adaptability of the framework to new domains while preserving alignment with human standards.

6 Limitations

The proposed framework, while offering a structured approach to LLM-generated content evaluation, faces several significant limitations that warrant careful consideration. Primary among these is the dataset's fundamental limitations, including the absence of explicit user persona specifications and temporal context, which are essential for accurately evaluating personalized or time-sensitive content. The research combines high-quality LLM-generated content with artificially degraded samples through semantic perturbations and linguistic modifications, which may not reflect realistic quality variations in real-world content. Additionally, the use of a high-performing LLM generator may introduce a positive quality bias, complicating objective evaluation and potentially skewing performance metrics.

A significant concern lies in the human annotation process, which introduces considerable noise into the system. The data shows extremely poor inter-annotator agreement, along with negative ICC values, indicating fundamental reliability issues with the ground truth labels. The reliance on single-evaluator annotations per sample introduces potential for individual bias and limits the reliability of baseline assessments, as reflected in the observed data quality issues.

The framework itself presents inherent constraints that affect its reliability. Evaluation outcomes remain sensitive to subjectivity, as checklist-based scoring, while standardized, can vary significantly across annotators. The framework's dependence on the capabilities and output stability of the underlying LLM introduces additional constraints, including susceptibility to API rate limits and model updates. Furthermore, the accuracy of overall content scores is contingent on the quality of initial checklist predictions, meaning early-stage errors can propagate through the evaluation pipeline.

7 Future Work

The advancement of AI-generated text evaluation requires several critical research directions that build upon our current checklist-driven framework. Our findings indicate the need for a more robust and comprehensive evaluation ecosystem that addresses current limitations while anticipating future challenges.

Primary focus areas include expanding content diversity and implementing multi-annotator protocols. The evaluation pipeline must adapt to diverse domains, including educational resources, commercial blogs, and technical documentation, while capturing domain-specific quality criteria. Multilingual adaptation represents another crucial development path, requiring culturally sensitive checklist items and cross-lingual evaluation [Doddapaneni et al. 2024] models to facilitate assessment across languages and regions, addressing the current gap in non-English evaluation.

Interactive and dynamic evaluation protocols present the next frontier, particularly in assessing content quality during multi-turn human-model interactions [Lee et al. 2024]. These protocols will provide deeper insights into model behavior and user experience over time. Additionally, enhancing human annotation practices through improved training and consensus-building mechanisms will strengthen ground-truth label validity. The integration of multimodal evaluation, particularly in domains requiring visual elements, necessitates new metrics for text-image coherence and joint semantic alignment [Yang et al. 2024].

Further validation requires comparison against stronger learned judges such as Prometheus and L-Eval, along with testing on larger, human-written and multilingual datasets. Personalization strategies must also be developed to incorporate user-specific context, including demographics and interaction history. These improvements, combined with robust bias mitigation strategies and hybrid human-AI pipelines, will create a more comprehensive and adaptive evaluation framework that better serves real-world AI applications.

8 Impact Statement

The increasing reliance on AI for long-form content generation by creators and organizations presents substantial challenges in balancing scalable quality evaluation with human-aligned standards. Existing solutions are hampered by three critical gaps: (1) inefficient manual assessment processes that elevate costs and limit the granularity of actionable feedback, (2) unreliable detection of LLM-generated biases-including factual inaccuracies and cultural stereotypes-and (3) systemic inequities that advantage users with access to advanced LLM APIs over resource-constrained creators. These limitations risk propagating low-quality or biased outputs, eroding user trust, amplifying misinformation, and introducing fragility into content pipelines due to API dependencies.

The HALF-Eval framework directly addresses these challenges by automating human-aligned quality assessment through structured checklists and regression modeling. This approach reduces evaluation costs, improves content depth and coherence via targeted feedback, and provides a scalable, reproducible methodology for diverse content types. However, the framework also introduces new risks. Its current focus on English-language content and checklist-driven metrics may inadvertently marginalize non-English creators and homogenize content styles, reinforcing language and cultural biases present in LLM training data. Dependency on LLM APIs further risks perpetuating Western-centric norms and exacerbating access inequities.

We have documented our datasets, models, and evaluation protocols to facilitate transparency, reproducibility, and external scrutiny. We acknowledge the importance of ongoing bias mitigation, inclusive evaluation design, and responsible artifact release to minimize potential harms. To the best of our knowledge, this work does not pose foreseeable risks of societal harm or misuse; rather, it aims to advance responsible, equitable, and human-aligned evaluation practices in generative AI. We encourage future research to address language inclusivity, personalization, and cross-cultural adaptation to further broaden the positive impact of automated content evaluation systems.

References

2024. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. arXiv:2307.11088 [cs.CL] https://arxiv.org/abs/2307.11088

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024a. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. arXiv:2308.14508 [cs.CL] https://arxiv.org/abs/2308.14508

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs. arXiv:2408.07055 [cs.CL] https://arxiv.org/abs/2408.07055

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*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909/

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000104 arXiv:https://doi.org/10.1177/001316446002000104

Jack Conklin. 2005. *Educational Horizons* 83, 3 (2005), 154–159. http://www.jstor.org/stable/42926529

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, and Mitesh M. Khapra. 2024. Cross-Lingual Auto Evaluation for Assessing Multilingual LLMs. arXiv:2410.13394 [cs.CL] https://arxiv.org/abs/2410.13394

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A Comprehensive Benchmark for Evaluating Long Text Modeling Capacities of Large Language Models. arXiv:2309.13345 [cs.CL] https://arxiv.org/abs/2309.13345

Shahriar Golchin and Mihai Surdeanu. 2024. Time Travel in LLMs: Tracing Data Contamination in Large Language Models. arXiv:2308.08493 [cs.CL] https://arxiv.org/abs/2308.08493

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2022. Efficient Long-Text Understanding with Short-Text Models. arXiv:2208.00748 [cs.CL] https://arxiv.org/abs/2208.00748

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. arXiv:2310.08491 [cs.CL] https://arxiv.org/abs/2310.08491

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models. arXiv:2406.05761 [cs.CL] https://arxiv.org/abs/2406.05761

Timothy Lau and Alex Biedermann. 2020. Assessing AI Output in Legal Decision-Making with Nearest Neighbors. *Penn State Law Review* 124 (2020), 609–655. Available at SSRN: https://ssrn.com/abstract=3459870.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2024. Evaluating Human-Language Model Interaction. arXiv:2212.09746 [cs.CL] https://arxiv.org/abs/2212.09746

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv:2305.14251 [cs.CL] https://arxiv.org/abs/2305.14251

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

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 (ACL)*. Philadelphia, 311–318.

John Pavlik. 2023. Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism and Mass Communication Educator* 78 (01 2023), 107769582211495. https://doi.org/10.1177/10776958221149577

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. 2024. HelloBench: Evaluating Long Text Generation Capabilities of Large Language Models. arXiv:2409.16191 [cs.CL] https://arxiv.org/abs/2409.16191

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs.CL] https://arxiv.org/abs/1908.10084

Domenic Rosati, Robie Gonzales, Jinkun Chen, Xuemin Yu, Melis Erkan, Yahya Kayani, Satya Deepika Chavatapalli, Frank Rudzicz, and Hassan Sajjad. 2024. Long-form evaluation of model editing. arXiv:2402.09394 [cs.CL] https://arxiv.org/abs/2402.09394

Aymeric Roucher. 2025. *Using LLM-as-a-judge for an automated and versatile evaluation.* https://huggingface.co/learn/cookbook/en/llm_judge Accessed April 30, 2025.

Amazon Web Services. 2025. *Amazon Nova Foundation Models.* https://aws.amazon.com/nova Accessed April 30, 2025.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding. arXiv:2305.14196 [cs.CL] https://arxiv.org/abs/2305.14196

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison Over Long Language Sequences. arXiv:2201.03533 [cs.CL] https://arxiv.org/abs/2201.03533

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. arXiv:2402.14207 [cs.CL] https://arxiv.org/abs/2402.14207

P. E. Shrout and J. L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 86, 2 (March 1979), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. PROXYQA: An Alternative Framework for Evaluating Long-Form Text Generation with Large Language Models. arXiv:2401.15042 [cs.CL] https://arxiv.org/abs/2401.15042

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL] https://arxiv.org/abs/1706.03762

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2025. CollabStory: Multi-LLM Collaborative Story Generation and Authorship Analysis. arXiv:2406.12665 [cs.CL] https://arxiv.org/abs/2406.12665

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv:2303.04048 [cs.CL] https://arxiv.org/abs/2303.04048

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. arXiv:2403.18802 [cs.CL] https://arxiv.org/abs/2403.18802

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal G. Gajera, Shreeyash Mukul Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL – A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models. arXiv:2305.05050 [cs.CL] https://arxiv.org/abs/2305.05050

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A Critical Evaluation of Evaluations for Long-form Question Answering. arXiv:2305.18201 [cs.CL] https://arxiv.org/abs/2305.18201

Jheng-Hong Yang, Carlos Lassance, Rafael S. Rezende, Krishna Srinivasan, Stéphane Clinchant, and Jimmy Lin. 2024. Retrieval Evaluation for Long-Form and Knowledge-Intensive Image–Text Article Composition. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, Lucie Lucie-Aimée, Angela Fan, Tajuddeen Gwadabe, Isaac Johnson, Fabio Petroni, and Daniel van Strien (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 36–45. https://doi.org/10.18653/v1/2024.wikinlp-1.9

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving Long Story Coherence With Detailed Outline Control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3378–3465. https://doi.org/10.18653/v1/2023.acl-long.190

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text. arXiv:2305.13304 [cs.CL] https://arxiv.org/abs/2305.13304

.1 Figures

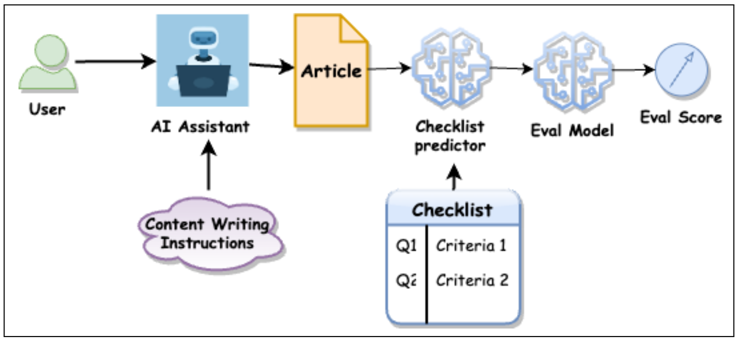

Figure 3: HALF-EVAL: Scoring diagram

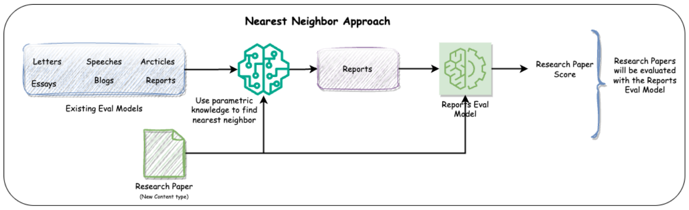

Figure 4: HALF-EVAL: Scaled Evaluation - Nearest Neighbor Mapping

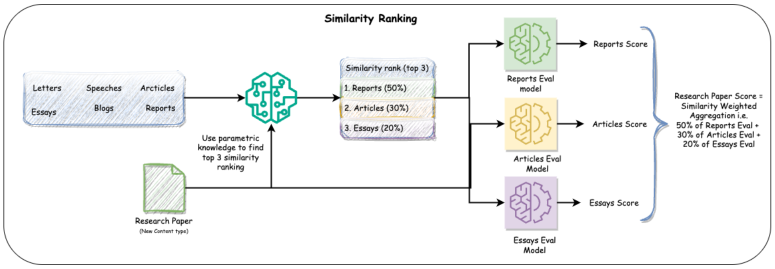

Figure 5: HALF-EVAL: Scaled Evaluation - Similarity Ranking

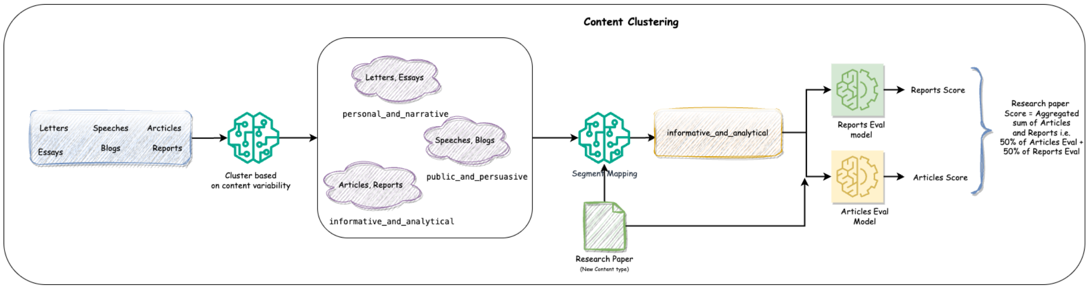

Figure 6: HALF-EVAL: Scaled Evaluation - Content Clustering

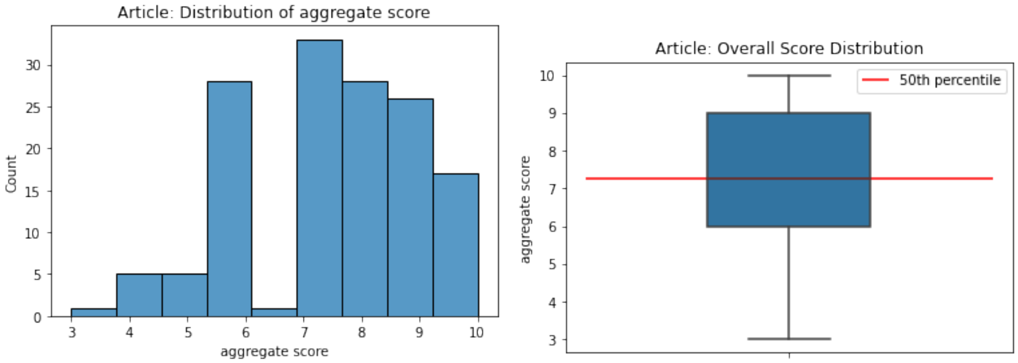

Figure 7: Article: Overall Score Distribution

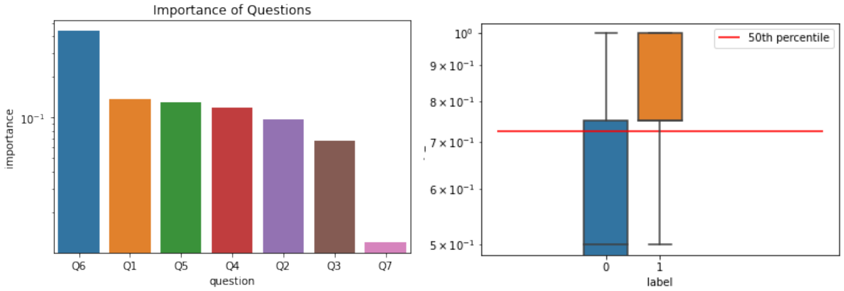

Figure 8: Article-Feature importance[Left], Label-Q6-Score Distribution[right]

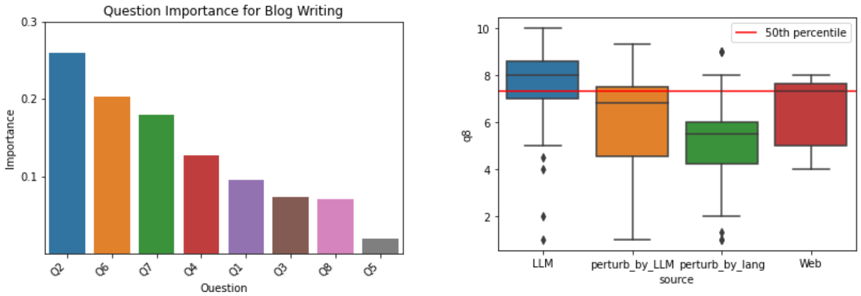

Figure 9: Blogs: Checklist Importance[Left], Score Distribution[Right]

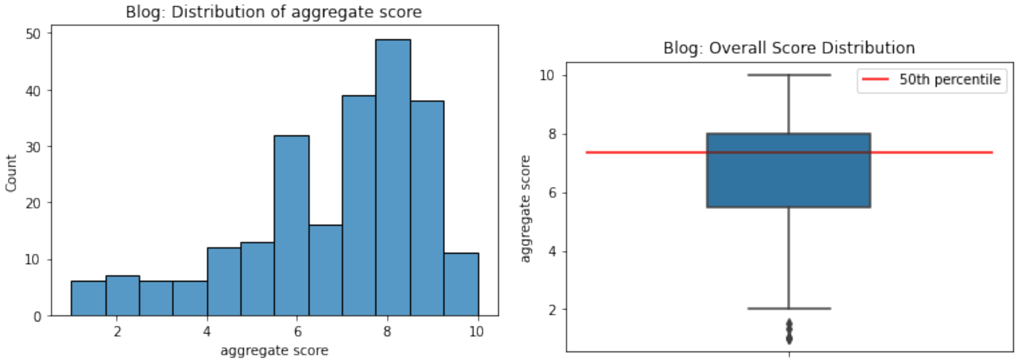

Figure 10: Blog: Overall Score Distribution

.2 Checklist Questions

Table 2: Article - Checklist Question

Category	Checklist
Creativity and Originality	Does the generated article fully align with the writing prompt, thoroughly and creatively respond to its content, and consistently capture and enhance its intended theme, tone, nuances, and deeper meanings throughout, while adding depth and originality to the prompt's concept?
Interest and Impact	Is the generated article consistently engaging, highly original, and novel, compelling readers to continue reading with a strong desire for more due to its captivating and intriguing narrative? Is the generated content so highly persuasive, with compelling arguments, credible evidence, and convincing reasoning throughout, that after reading the entire content, you are unable to find any points to refute the arguments presented?
Coherence and Consistency	Does the generated article comprehensively address the thesis, present thoroughly developed arguments with substantial evidence, conclude in a convincing manner, and consistently maintain rigorous logical coherence and alignment of viewpoints throughout?
Relevance	Does the generated text aligns with the given prompt ? Is the generated content not only sufficiently long and complete but also thoroughly detailed, ensuring each argument is extensively explained and supported by comprehensive evidence?"

Table 3: Blog - Checklist Question

Category	Checklist
Creativity and Originality	Does the content provide a unique perspective?
Interest and Impact	Is there depth, or is it too generic? Does the writing have a clear, engaging tone?
Coherence and Consistency	Are sentences concise and naturally structured?
Relevance	Is storytelling used effectively? Does the content fully answer the prompt and stay on-topic?

.3 Table

Table 4: Article: Feature Selection

Model	Train Features	Output	MAE
LinearRegression	Individual Checklist Scores	Over-All-Score	0.821
LinearRegression	Individual Checklist Scores + prompt +text	Over-All-Score	1.521
LinearRegression	Prompt + Text	Over-All-Score	1.691
LinearRegression	Text + Individual Checklist Scores	Over-All-Score	1.847
LinearRegression	Prompt + Individual Checklist Scores	Over-All-Score	1.911

Table 5: Article: Model Selection

Model	Features	Output	MAE
RandomForestRegressor	Individual Checklist Scores	Over-All-Score	0.659
GradientBoostingRegressor	Individual Checklist Scores	Over-All-Score	0.665
Ridge	Individual Checklist Scores	Over-All-Score	0.815
LinearRegression	Individual Checklist Scores	Over-All-Score	0.821
DecisionTreeRegressor	Individual Checklist Scores	Over-All-Score	0.826
Lasso	Individual Checklist Scores	Over-All-Score	1.227
ElasticNet	Individual Checklist Scores	Over-All-Score	1.227

Table 6: Blog: Feature Selection

Model	Train Features	Output	MAE
LinearRegression	Individual Checklist Scores	Over-All Score	0.08
LinearRegression	Text + Individual Checklist Scores	Over-All Score	0.20
LinearRegression	prompt + Individual Checklist Scores	Over-All Score	0.21
LinearRegression	prompt + text + Individual Checklist Scores	Over-All Score	0.27

Table 7: Blog: Model Selection

Model	Train Features	Output	MAE
LinearRegression	Individual Checklist Scores	Over-All Score	0.075
RandomForestRegressor	Individual Checklist Scores	Over-All Score	0.075
FeedForwardNeuralNetwork	Individual Checklist Scores	Over-All Score	0.076
GradientBoosterRegression	Individual Checklist Scores	Over-All Score	0.08

.4 Annotator Agreement

Table 8: Annotator Agreement

Content Type	k-cappa	ICC1
Article	0.296	-0.997
Blog	0.162	-0.495

Table 9: HALF-EVAL-KPI: Evaluation KPIs with ground truth. Essays did not have the KPI level ground truth.

	Article			Blog			Essay		
Ground truth vs HALF-EVAL	MAE	p-val	95% CI	MAE	p-val	95% CI	MAE	p-val	95% CI
Creativity and Originality	2.06	0.24	0.74	1.17	0.11	0.63	-	-	-
Interest and Impact	3.5	0.03	0.75	1.71	0.0	1.02	-	-	-
Coherence and Consistency	3.93	0.03	0.85	1.77	0.28	1.27	-	-	-
Relevance	2.14	0.46	0.5	2.16	0.03	0.98	-	-	-
Aggregate Score-HALF-EVAL	1.43	0.1	0.33	2.04	0.04	1.13	3.14	0.71	0.25
Aggregate Score-LLMJ	1.28	0.33	0.29	2.01	0.02	1.12	4.29	0.0	0.46

Table 10: Scaling Result: Essay

	Essay		
Metric	**MAE**	**p-val**	**95% CI**
Aggregate Score-Similarity	3.14	0.71	0.25
Aggregate Score-Nearest Neighbor	3.21	0.31	0.25
Aggregate Score-Cluster	3.21	0.31	0.25
Aggregate Score-LLMJ	4.29	0.0	0.46

```
You are tasked with adding perturbation to the given blog content
according to specific guidelines. Your goal is to modify the text
while maintaining a similar word count. Here's what you need to do:

1. Apply the following perturbation guidelines to the content:
- Add random words
- Add meaningless words
- Change words to synonyms
- Remove SEO-optimized keywords
- Remove insights and details
- Remove references and citations
- Remove introduction
- Remove summary and conclusion
- Remove quantitative data
- Change the order of sentences
- Paraphrase heavily
- Change the syntactic structure
 2. When applying these perturbations:
 - Ensure that the overall word count remains similar to the
   original content
 - Apply the perturbations randomly throughout the text
 - Make sure the resulting text is still readable, albeit less
     coherent and informative than the original
 3. After applying the perturbations, provide the modified content
    as your output. Do not include any explanations, comments, or
    additional text.
 4. Important: Your response should contain ONLY the perturbed
    content. Do not include any other text, explanations, or
    formatting outside of the content itself.

Begin the perturbation process for the following content and
provide only the modified content as your output.

<content> {{CONTENT}} </content>
```

.5 Perturbation Prompt

.6 Checklist predictor: LLM-as-a-Judge

```
You are an expert creative content evaluator. Your task is to
analyze the given {{content_type}} based on the provided
checklist of questions and generate a detailed evaluation
report in JSON format.

Input Parameters:
1. User prompt: {{user_prompt}}
2. Content text: [{{content_text}}]
3. List of questions: [{{list_of_questions}}]

Please evaluate the {{content_type}} by:
1. Answering each question from the checklist
2. Providing a Yes/No for each criterion
3. Providing a score (0-10) for each criterion
4. Giving a brief reason for each score
5. Calculating an overall score as average of all
   criteria (0-10)

Present your evaluation in the following JSON format:

{ "{{COL_PROMPT}}": "The original user prompt",
"q1": "Original question1",
"q1_response": "No",
"q1_score": 0.0,
"q1_reason": "Reason for the score",
"q2": "Original Question2",
"q2_response": "Yes",
"q2_score": 10.0,
"q2_reason": "Reason for the score",
"{{COL_AGGREGATE_SCORE}}": 0,
"summary": "A brief summary of the evaluation" }

Guidelines for Yes/No:
- No: When content text does not meet the criterion
- Yes: When content text meets the criterion

Guidelines for scoring:
- 0-3: Poor/Unsatisfactory
- 4-6: Average/Needs Improvement
- 7-8: Good
- 9-10: Excellent

Ensure that:
1. Each score is justified with a clear,
  specific reason
2. The evaluation is objective and based on the
{{content_type}}'s content
3. Recommendations are actionable and specific
4. The general feedback summarizes the key strengths
and areas for improvement

Please provide your evaluation based on these parameters.
DO NOT GIVE ANYTHING ELSE OTHER THAN THE JSON OUTPUT
```