

SCORE: Story Coherence and Retrieval Enhancement for AI Narratives

Qiang Yi^{1*}, Yangfan He^{2*}, Jianhui Wang^{3*}, Xinyuan Song⁴, Shiyao Qian⁵, Xinhang Yuan⁶,
Li Sun⁷, Yi Xin⁸, Jingqun Tang⁹, Keqin Li¹⁰, Kuan Lu¹¹, Menghao Huo¹², Jiaqi Chen¹⁰, Tianyu Shi^{5†}

¹University of California, Berkeley, ²University of Minnesota – Twin Cities,

³University of Electronic Science and Technology of China, ⁴Emory University,

⁵University of Toronto, ⁶Washington University, Saint Louis, ⁷Amazon,

⁸Nanjing University, ⁹ByteDance Inc., ¹⁰Independent Researcher,

¹¹Cornell University, ¹²Santa Clara University

Project Page: <https://jianhuiwemi.github.io/SCORE>

Abstract—Large Language Models (LLMs) can generate creative and engaging narratives from user-specified input, but maintaining coherence and emotional depth throughout these AI-generated stories remains a challenge. In this work, we propose SCORE, a framework for Story Coherence and Retrieval Enhancement, designed to detect and resolve narrative inconsistencies. By tracking key item statuses and generating episode summaries, SCORE uses a Retrieval-Augmented Generation (RAG) approach, incorporating TF-IDF [1] and cosine similarity [2] to identify related episodes and enhance the overall story structure. Results from testing multiple LLM-generated stories demonstrate that SCORE significantly improves the consistency and stability of narrative coherence compared to baseline GPT models, providing a more robust method for evaluating and refining AI-generated narratives.

Index Terms—Collaborative Multi-Agent Framework, Language Model Fine-Tuning, Reinforcement Learning from Human Feedback

I. INTRODUCTION

Deep learning has transformed multiple domains including NLP, time series analysis and computer vision [3]–[8]. Large Language Models (LLMs) have demonstrated significant capabilities in generating long-form narratives, such as serialized stories or novels, by leveraging large-scale architectures and vast amounts of training data [9]. However, maintaining narrative consistency over extended texts, especially in terms of character development and emotional coherence, remains a major challenge [10]. For instance, [11] pointed out that achieving thematic consistency and managing dynamic plot states is crucial for maintaining the logical flow of a story. In practice, LLMs often struggle with inconsistencies when characters or key plot items reappear without proper explanation, disrupting the overall narrative structure.

Similarly, [12] highlight the difficulties in managing multimodal elements within long-form narratives, noting that inconsistencies in character behavior or emotional tone can negatively impact reader engagement. These challenges indicate a need for more structured approaches in narrative generation

that can better manage character arcs, plot developments, and emotional progression throughout the story.

In addition, recent works have highlighted the importance of memory mechanisms in LLM-based agents. [13] conducted a comprehensive survey on these mechanisms, identifying effective memory designs that help mitigate inconsistencies in narrative development—a challenge common to both interactive agents and narrative generation tasks. Additionally, [14] introduced generative agents that simulate human-like behavior using memory modules. These agents track the state of a wide array of interactable objects in a sandbox environment, ensuring consistent reasoning and enabling the smooth functioning of a simulated society. These researches inspired the design of our new framework.

In this work, we build upon recent advancements in Retrieval-Augmented Generation (RAG) [15], which dynamically incorporates relevant context to enhance narrative coherence. Expanding on these developments, we propose SCORE, a framework designed to evaluate three critical aspects of long-form narrative generation: character consistency, emotional coherence, and logical tracking of key plot elements. Our key contributions are:

- We introduce SCORE, an LLM-based evaluation framework that detects narrative inconsistencies in AI-generated stories.
- We incorporate a Retrieval-Augmented Generation (RAG) approach, utilizing episode-level summaries and key item tracking to improve narrative coherence.
- We demonstrate enhanced story consistency and emotional depth by integrating sentiment analysis and similarity-based episode retrieval.
- We outperform baseline GPT model [16] in detecting continuity errors and maintaining overall narrative coherence.

II. METHOD

Our proposed method, SCORE, consists of three main components: (1) an LLM-based evaluation framework to assess the coherence of key story elements, (2) automatic generation of episode summaries to track plot development, and (3) a

*Equal contribution.

†Corresponding author.

retrieval-augmented generation (RAG) approach that integrates the first two components, enabling enhanced user interaction and ensuring narrative consistency.

As the framework is intended solely for academic research purposes, its use is consistent with the original access conditions of all incorporated tools and data sources.

A. Continuity Analysis and Key Item Status Correction

By extracting key parts of GPT-4’s analysis, we identify instances where an item reappears in later episodes after being marked as lost or destroyed, without narrative explanation. Let $S_i(t)$ represent the state of item i at time t , where $S_i(t) \in \{\text{active, lost, destroyed}\}$. If item i reappears at time t_k with $S_i(t_k) = \text{active}$ after being previously marked as $S_i(t_{k-1}) \in \{\text{lost, destroyed}\}$, we flag this as a continuity error. To maintain consistency, the state remains $S_i(t_{k-1})$, avoiding an incorrect update. This approach systematically corrects discrepancies in item states, ensuring that narrative continuity is preserved by preventing erroneous state transitions.

B. Key Item Interaction Analysis

For each episode, we conduct a thorough evaluation by summarizing key plot points, character actions, and tracking interactions with important items. Let $A_c(t)$ represent the actions of character c at time t , and let $I_i(t)$ denote interactions with key item i . The model generates summaries that encapsulate essential elements, including $A_c(t)$ (character actions), relationships, and emotional changes across the episode. It then analyzes the specific interactions $I_i(t)$ between characters and key items, documenting these for further analysis. This step aggregates relevant narrative information—combining episode summaries, key item interactions $I_i(t)$, and character actions $A_c(t)$ —to facilitate more precise future retrieval. The approach simplifies subsequent analysis of plot and item continuity, reducing redundancy and improving efficiency.

C. Similarity-Based Episode Evaluation and Sentiment Analysis

We integrate similarity-based retrieval and sentiment analysis to improve episode evaluation and answer complex queries. It begins by loading summaries, full episode content, and key item states from structured JSON files. The content is segmented into smaller chunks using a text segmenter and embedded into a vector space model using FAISS [17] and OpenAI embeddings. This vector space enables efficient retrieval of similar episodes for user queries or specific episode analysis.

For evaluation, the system retrieves relevant past episodes by calculating cosine similarity scores between the current episode or query and all other episodes in the vector space. Let $S(e_c, e_p)$ represent the similarity score between the current episode e_c and a past episode e_p . The top N episodes with the highest $S(e_c, e_p)$ scores are retrieved for further analysis, providing a relevant summary of episodes for evaluation or answering questions.

Sentiment analysis is then applied to both the current and retrieved episodes. A sentiment score $\sigma(e)$, ranging from 0 to 1,

is assigned to each episode e by GPT-4, reflecting its emotional tone. These scores help refine the selection by ensuring both text similarity and sentiment consistency are considered, thus preventing errors from large sentiment discrepancies.

Finally, the LLM processes the retrieved episode summaries and content to generate a detailed evaluation. The focus is on narrative aspects such as character consistency, plot progression, emotional authenticity, and key item continuity. This ensures the narrative remains coherent, with any discrepancies flagged and corrected.

III. EXPERIMENTS

To evaluate the effectiveness of SCORE, we conducted experiments on stories generated by large language models (LLMs). These experiments assessed the framework’s ability to maintain narrative coherence, detect continuity errors, and ensure emotional consistency throughout episodic storytelling.

A. Dataset Preparation

We collected a diverse set of stories generated by various GPT models, covering a wide range of genres such as science fiction, drama, and fantasy. Each story was divided into multiple episodes, with each episode containing the raw outputs generated by the language models, such as dialogues between characters, descriptions of events, and narrative elements. This variety in genres and structures allowed us to thoroughly evaluate our framework’s performance across different storytelling formats.

B. Baselines

We compared our proposed framework to three baselines: GPT-4, GPT-4o, and GPT-4o-mini. In these cases, we used these models directly without integrating our key item tracking, continuity analysis, or retrieval-augmented generation (RAG) mechanisms. We used the same LLM-generated stories to evaluate different models. For all baselines, we measured their ability to evaluate episodes independently and answer complex questions, without deliberately guiding them through the details of the story.

C. Evaluation Process

For each episode in the dataset, the evaluation process involved several key steps. First, we conducted an initial evaluation by directly uploading the files to ChatGPT, testing if GPT could assess the story correctly without any prompts. We used GPT-4o-mini, GPT-4o, and GPT-4 for narrative evaluation.

Next, we configured the LLM for a more detailed evaluation, using our preprocessed files to ensure accurate tracking of key items across episodes. We also employed the Retrieval-Augmented Generation (RAG) framework to retrieve episodes similar to the one being evaluated. By using FAISS to calculate cosine similarity scores between episodes, the system was able to retrieve episodes that were semantically and emotionally aligned with the current one, providing additional context. This enriched context was then used to construct GPT prompts, enabling a more thorough evaluation of the episode.

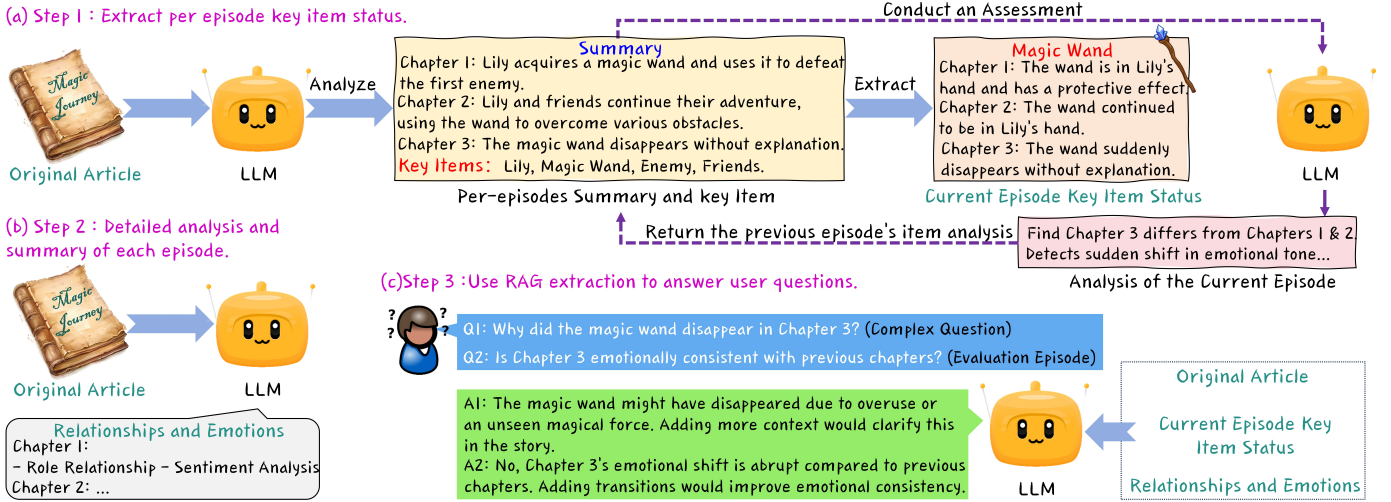


Fig. 1: The SCORE framework for improving AI-generated story coherence. (a) Extracts key item statuses per episode. (b) Conducts detailed analysis and summaries of each episode. (c) Uses RAG to answer user queries and resolve narrative inconsistencies.

TABLE I: Performance comparison of models with and without SCORE

Model	Consistency	Coherence	Item Status	Complex Question
GPT-4o-mini + SCORE	78.2 82.6 ($\uparrow 4.4$)	76.7 77.5 ($\uparrow 0.8$)	0 80.5 ($\uparrow 80.5$)	24.56 63.0 ($\uparrow 38.44$)
GPT-4o + SCORE	86.78 88.68 ($\uparrow 1.9$)	82.21 89.91 ($\uparrow 7.7$)	0 96 ($\uparrow 96$)	76.32 88.75 ($\uparrow 12$)
GPT-4 + SCORE	83.21 85.61 ($\uparrow 2.4$)	84.32 86.9 ($\uparrow 2.58$)	0 98 ($\uparrow 98$)	82.34 89.45 ($\uparrow 7.11$)

D. Metrics

We evaluated the framework based on several key metrics. Narrative coherence was assessed by examining how well the framework maintains the logic of the story. Specifically, we tracked the consistency of character behavior and plot development across episodes to ensure that continuity errors (especially those involving critical content) were detected. Finally, we scored the story evaluations obtained by the two methods to measure the stability of the framework.

E. Results

Our experiments demonstrated that the proposed framework significantly improved the detection of narrative inconsistencies. Evaluations using the framework were able to more accurately detect inconsistencies in character actions or plot progression. The retrieval-augmented generation process helped GPT better filter irrelevant information, understand the current story context, and improved its ability to detect narrative continuity across multiple episodes. When quantitatively compared to baseline methods, such as using GPT model alone, the proposed framework showed substantial improvements in evaluation accuracy.

IV. LIMITATIONS

While our SCORE framework offers significant improvements, it also presents several limitations:

- **Dependence on Retrieval Accuracy:** The effectiveness of SCORE relies on accurately retrieving relevant episodes. Errors in similarity calculation or sentiment analysis may lead to the exclusion of important contextual information, affecting the overall evaluation quality.
- **Challenges in Capturing Emotional Nuances:** Sentiment analysis may not fully capture the complexity of emotions, potentially missing subtle shifts or deeper emotional layers within the narrative.
- **Resource Intensive:** Despite optimizations, SCORE still requires considerable computational resources for similarity calculations and sentiment analysis, which can be limiting for large datasets.
- **Generalizability:** SCORE has been primarily tested on GPT-4-generated stories, and its performance on narratives from other models or with different structures remains an open question.

V. CONCLUSION

We introduced SCORE, a novel LLM-based evaluation framework aimed at improving the coherence and emotional consistency of AI-generated stories. By incorporating Retrieval-Augmented Generation (RAG), TF-IDF, cosine similarity, and sentiment analysis, our method retrieves contextually relevant information, enabling more accurate and detailed assessments.

Experimental results demonstrate that SCORE outperforms traditional methods by effectively identifying continuity errors, maintaining narrative coherence, and providing valuable insights for episodic storytelling. However, limitations such as dependency on retrieval accuracy and computational demands present opportunities for future improvement.

In future work, we aim to optimize retrieval accuracy, refine emotional analysis, and expand the framework’s applicability to a broader range of LLM-generated stories to further validate its adaptability and generalizability.

REFERENCES

- [1] Juan Ramos et al., “Using tf-idf to determine word relevance in document queries,” 2003.
- [2] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al., “Semantic cosine similarity,” 2012.
- [3] Xiangfei Qiu, Xiuwen Li, Ruiyang Pang, Zhicheng Pan, Xingjian Wu, Liu Yang, Jilin Hu, Yang Shu, Xuesong Lu, Chengcheng Yang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, and Bin Yang, “Easytime: Time series forecasting made easy,” in *ICDE*, 2025.
- [4] Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang, “Duet: Dual clustering enhanced multivariate time series forecasting,” in *SIGKDD*, 2025, pp. 1185–1196.
- [5] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang, “Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods,” in *Proc. VLDB Endow.*, 2024, pp. 2363–2377.
- [6] Shutao Li, Bin Li, Bin Sun, and Yixuan Weng, “Towards visual-prompt temporal answer grounding in instructional video,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 12, pp. 8836–8853, 2024.
- [7] Bin Li and Hanjun Deng, “Bilateral personalized dialogue generation with contrastive learning,” *Soft Computing*, vol. 27, no. 6, pp. 3115–3132, 2023.
- [8] Bin Li, Bin Sun, Shutao Li, Encheng Chen, Hongru Liu, Yixuan Weng, Yongping Bai, and Meiling Hu, “Distinct but correct: generating diversified and entity-revised medical response,” *Science China Information Sciences*, vol. 67, no. 3, pp. 132106, 2024.
- [9] Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie, “Rolecraft-glm: Advancing personalized role-playing in large language models,” in *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 2024, pp. 1–9.
- [10] Dan P McAdams, “The problem of narrative coherence,” *Journal of constructivist psychology*, vol. 19, no. 2, pp. 109–125, 2006.
- [11] Aisha Khatun and Daniel G Brown, “Assessing language models’ worldview for fiction generation,” *arXiv preprint arXiv:2408.07904*, 2024.
- [12] Danyang Liu, Mirella Lapata, and Frank Keller, “Generating visual stories with grounded and coreferent characters,” *arXiv preprint arXiv:2409.13555*, 2024.
- [13] Zeyu Zhang, Xiaohu Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen, “A survey on the memory mechanism of large language model based agents,” *arXiv preprint arXiv:2404.13501*, 2024.
- [14] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein, “Generative agents: Interactive simulacra of human behavior,” 2023.
- [15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K ttler, Mike Lewis, Wen-tau Yih, Tim Rock schel, et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, “Language models are unsupervised multitask learners,” 2019, Technical report.
- [17] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilv s, Pierre-Emmanuel Mazar , Maria Lomeli, Lucas Hosseini, and Herv  J gou, “The faiss library,” 2024.

APPENDIX

Table 3 to 6 display the results from both using the GPT API and directly utilizing the GPT playground to analyze the article. In the GPT playground, the full episode script is uploaded, and the prompt “Please evaluate episode n” is used to generate baseline results. When calling the GPT API, we upload the full episode and processed files, employ the OpenAI API combined with our framework to answer user queries, focusing primarily on evaluating the coherence of each episode. As noted earlier, the evaluation process has been optimized, resulting in a 90% improvement in analysis accuracy.

A. Proof of the Continuity Analysis of Critical Item State Correction

Assume a state space \mathcal{M} containing states such as “active” and “destroyed” (with “lost” similarly defined). The system satisfies the Markov property:

$$P(S(t_k) = s \mid S(t_{k-1}) = s_{k-1}, \dots, S(t_0) = s_0) = P(S(t_k) = s \mid S(t_{k-1}) = s_{k-1}). \quad (1)$$

Denoting destroyed and lost as absorbing states, This mathematically means:

$$P(S(t_k) = \text{destroyed} \mid S(t_{k1}) = \text{destroyed}) = 1. \quad (2)$$

Thus, no transition from destroyed or lost back to active or any non-terminal state is allowed.

We define the narrative entropy function of a state sequence by [10]. Let

$$E(S_0, S_1, \dots, S_T) = \sum_{k=0}^{T-1} \Delta E(S_k, S_{k+1}), \quad (3)$$

be the total narrative entropy over the sequence, where $\Delta E(S_k, S_{k+1})$ represents the increase in entropy when transitioning from state S_k to state S_{k+1} . It has $\Delta E(S_k, S_{k+1}) = -\ln(P(S_{k+1} \mid S_k))$.

Then, the total narrative entropy over the sequence $\{S_0, S_1, \dots, S_T\}$ is:

$$E(S_0, S_1, \dots, S_T) = -\sum_{k=0}^{T-1} \ln(P(S_{k+1} \mid S_k)). \quad (4)$$

Because narratives are expected to be coherent, the theory requires that the entropy increases are minimal (ideally zero) along the legal state path. That is, for all legal transitions $S_k \rightarrow S_{k+1}$, we have

$$\Delta E(S_k, S_{k+1}) = 0. \quad (5)$$

Now suppose, by way of contradiction, that an illegal transition is allowed. That is, there exists some k for which

$$P(S_{k+1} = \text{active} \mid S_k = \text{destroyed}) > 0. \quad (6)$$