AIStorySimilarity: Quantifying Story Similarity Using Narrative for Search, IP Infringement, and Guided Creativity

Jon Chun
Kenyon College
chunj@kenyon.edu

Abstract

Stories are central for interpreting experiences, communicating, and influencing each other via films, medical, media, and other narratives. Quantifying the similarity between stories has numerous applications including detecting IP infringement, detecting hallucinations, search/recommendation engines, and guiding human-AI collaborations. Despite this, traditional NLP text similarity metrics are limited to short text distance metrics like n-gram overlaps and embeddings. Larger texts require preprocessing with significant information loss through paraphrasing or multi-step decomposition. This paper introduces AIStorySimilarity, a novel benchmark to measure the semantic distance between long-text stories based on core structural elements drawn from narrative theory and script writing. Based on four narrative elements (characters, plot, setting, and themes) as well 31 sub-features within these, we use a SOTA LLM (gpt-3.5-turbo) to extract and evaluate the semantic similarity of a diverse set of major Hollywood movies. In addition, we compare human evaluation with story similarity scores computed three ways: extracting elements from film scripts before evaluation (Elements), directly evaluating entire scripts (Scripts), and extracting narrative elements from the parametric memory of SOTA LLMs without any provided scripts (GenAI). To the best of our knowledge, AIStorySimilarity is the first benchmark to measure long-text story similarity using a comprehensive approach to narrative theory. All code, data, and plot image files are available at https://github.com/jon-chun/AIStorySimiliarity.

1 Introduction

Stories and narrative are universally used by humans to communicate, interpret, store, and react to the world around them (Boyd, 2017) (Schreiner et al., 2017). When organized within a narrative framework, information can be more readily understood, stored, and recalled (Zdanovic et al., 2022). Beyond traditional fiction, researchers are now applying narrative theory to enhance medicine (Coret et al., 2018), law (Jiang et al., 2024b), business (Rees, 2020), and national identity rhetoric (Sweet and McCue-Enser, 2010). Narratives show immense potential for emotional persuasion (Lehnen, 2016) and, when used in combination with emotionally intelligent AI (Broekens et al., 2023), are classified as high risk by the EU AI Act (EU (Parliament) - and Jaume Duch Guillot, 2023).

A number of traditional NLP subtasks relate to stories and narratives, both for analysis and generation. Analysis is typically restricted to short-text lengths from approximately one sentence to several paragraphs at most (e.g. MoverScore, BERTScore, QAEval). NLP tasks include identifying sentiment, topics, characters, dialog, and events. Long texts can be analyzed with a sequential sliding-window of short-text substrings. This enables the extraction of distributed narrative elements from long texts including character social networks (Bost and Labatut, 2019), event timelines (Zhong and Cambria, 2023) or plot related information like diachronic-emotional arcs (Chun 2018) and narrative crux points (Elkins 2022).

Most traditional NLP techniques like sentiment classification, NER, and POS limit story analysis to relatively short texts. However, the introduction of the Transformer architecture (Vaswani et al., 2017) and rapid progress in LLM performance since the launch of ChatGPT (OpenAI, 2022) has revolutionized NLP. While smaller traditional models like BERT and BART can still be competitive for structured narrow tasks like NER (Paper with Code, 2024a) and POS (Papers with Code, 2024b), LLMs generally dominate the NLP leaderboards (Guo et al., 2023). More importantly, trained on trillions of tokens of language, LLMs have acquired a fluency, coherence, common-sense reasoning, expressiveness, and creativity with natural language that enable new, more complex and open-ended

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 161–177
November 15-16, 2024 ©2024 Association for Computational Linguistics

NLP tasks like human-level story generation (Xie et al., 2023) and analysis (Chun and Elkins 2023).

However, there are serious limitations to trying to understand long-text stories by using short-text NLP techniques over a sequence of sentences or paragraphs. Authors, readers, and IP lawyers generally evaluate stories at higher levels of abstractions that escape short-text decomposition techniques or suffer information loss in the process. Powerful narrative elements like character arcs, themes and complex plot devices are often latent, implied, and disseminated throughout the text and require a global unified perspective to identify, extract, and analyze. Narrative theory and screenwriting conventions provide conceptual frameworks for describing and capturing these essential structural elements inherent in stories.

Film studies, Narratology (Berhe et al., 2022) and script writing best practices (Mckee, 1997; Snyder, 2005; Truby, 2007) decompose narrative structures and elements into different narrative elements. Characters including relationships and motivations. Plot is the sequence of events in the story. Settings involve not only time and place, but other aspects like culture. Themes are central ideas and messages. Character arcs track the transformation of characters over the course of the story in response to events. Dialog collectively is the spoken words and interactions that reveal personality, relationships and advance the plot. Classification of narrative elements are flexible. A simpler framework could combine character, character arc and dialog into one broader concept of character. Arguably least intuitive, themes are the big ideas and messages that provide deeper meaning, emotional connection and purpose like good vs evil, life finds a way, or love endures.

A variety of NLG subfields try to leverage hallucination as a creativity control in story generation (Chieh-Yang et al., 2023), creative writing (Ippolito et al., 2022), and screenwriting (Mirowski et al., 2022). Text generation (CTG) is focused on controlling the creative process including more precisely directing the degree and type of hallucinations (Zhang et al., 2022). This could enhance human-AI interactions from better human-AI creativity collaboration to more engaging chatbots.

A relatively recent and small set of researchers have begun focusing on the positive value LLM hallucinations can bring in the form of creativity or 'confabulation' (Sui et al., 2024). This growing perspective warrants a survey of hallucination from a creative perspective (Jiang et al., 2024a), and new applications are being identified like contrastive dataset generation (Yao et al., 2023). The all rely upon upon semantic distance metrics.

The use case of quantifying intellectual property infringement of copyrighted works illustrates the concept of narrative 'similarity'. IP infringement upon written work like movie script involves two tests of 'substantial similarity'. The intrinsic test is an analysis of identifiable properties like character, plot points, and themes. The extrinsic test is a more subjective analysis of whether an "ordinary person" would recognize such similarities (Helfing, 2020). Unlike the high-profile NYTimes-OpenAI lawsuit claiming perfect word-for-word reproductions (Pope, 2024), most infringement cases have historically fallen in this gray zone of 'substantial similarity'. Many more cases may arise either accidentally or intentionally as generative AI becomes a mainstream content creation- and creative collaboration-tool. There is therefore a pressing need to formalize a semantic similarity metric for narratives. The main contributions of this paper are:

· AIStorySimilarity, the first narrative semantic similarity benchmark using a scoring rubric based on formal narrative structural elements. · Evaluation of three common comparison methodologies to measure the similarity between test and reference film narratives on a) parametric memory [GenAI], b) extracted narrative elements [Elements] and c) unprocessed scripts [Scripts]

· A benchmark with broad application for detecting IP infringement of copyrighted works, film/novel/narrative search and recommendation engines, detecting hallucinations, and guiding creativity with extensive reporting for human-in-the-loop explainability and verification.

2 Related Work

SemEval22 Task 8 evaluated the semantic distances between news stories in order to move to more complex semantic metrics. Many entrants used text representations like TF-IDF derived from traditional low-level syntax features (Jobanputra and Rodríguez, 2022), but others used features based on higher-level abstractions like narrative schemas and writing style (Chen et al., 2022). However, many of NLG evaluations using high-level abstractions like empathy and style (Shen et al., 2024) and the narrative theory of Labov and Waletzky (Levi et al.,

2022) focus on creating novel annotated training datasets (Chaturvedi et al., 2018). The 6th Annual Workshop on Narrative Extraction from Text (Campos et al., 2023) survey papers provide a contemporaneous overview of some of the more recent approaches to extracting narrative elements from text (Zhu et al., 2023).

Beyond AI text generation, SOTA LLMs like GPT3.5 and GPT4o are increasingly used as proxies for human evaluators in open-ended, reference-free NLG tasks (Li et al., 2024). They provide benefits of speed, scalability, and cost savings alongside increasingly human-level or better performance (Hada et al., 2023; Ke et al., 2024; Wang et al., 2023). This LLM-as-judge trend (Thakur et al., 2024) is evident in various NLP tasks, such as evaluating the quality of generated stories, assessing the effectiveness of adversarial attacks, and grading the comprehensibility of disordered speech transcriptions (Chiang and yi Lee, 2023; Tomanek et al., 2024). For instance, the MT-Bench framework demonstrates a strong 80% agreement between LLM evaluations and human judgments in assessing model performance (Zheng et al., 2023).

For semantic text similarity, LLMs are shown to be more aligned with humans than any other metric (Aynetdinov and Akbik, 2024). However, precautions must be taken to avoid biases like a model's preference for evaluating its own generated content (Chhun et al., 2024). Moreover, challenges remain in areas of trust and safety (Reiter, 2024) and problems exist with human evaluations themselves (Elangovan et al., 2024; Gao et al., 2024). Despite these limitations (Bavaresco et al., 2024), LLMs show promise in augmenting and even replacing certain types of human evaluations given continual advances in AI.

At higher levels of abstraction, a variety of research areas relate to text similarity. This includes subfields that rely upon structural elements for automatic story generation (ASG) or for automated essay scoring (AES). Traditionally, these fields have used a combination of human evaluators, human-annotated references, and more general NLP metrics like coherence (Guan et al., 2021). In addition, more formal structural approaches generate or evaluate more diffuse global features like narrative frameworks (Wang et al., 2022), readability (coherence, fluency, simplicity), and adequacy (faithfulness, informativeness) (Hu et al., 2024). Emphasis on story similarity between reference and test works relate to plagiarism detection, intellectual property infringement, movie recommendation and search engines, hallucination detection (Huang et al., 2023; Ye et al., 2023) and measuring creativity in derivative works. AIStorySimilarity leverages an abstract structural approach using narrative theory with similarity metrics using SOTA LLMs.

3 Methods

3.1 Dataset

To provide a reference to assess the accuracy of similarity scores and relative rankings, a human expert selected a dataset of 9 popular Hollywood films they ranked as shown in Table 1. "Raiders of the Lost Arc", a 1981 summer hit, was selected as the reference film and 8 other test films were selected in order of decreasing similarity. This included a. the 1984 and 1989 Indiana Jones sequel films, b. three other adventure genre films with historical artifact themes, and c. three very different non-adventure genre films (romantic drama, black comedy, and musical). All scripts are ingested as plain text complete with character name, dialog, scene headings, action, and other annotations where available (see ./data/film_scripts_txt).

Sim.	Genre	Name	Year	Rank
ref	Adventure	Raiders of the Lost Ark	1981	-
1	Sequel #1	Indiana Jones and the Temple of Doom	1984	2
2	Sequel #2	Indiana Jone and the Last Crusade	1989	2
3	Adventure	National Treasure	2004	10
4	Adventure	Laura Croft Tomb Raider	2001	14
5	Adventure	The Mummy	1999	8
6	Romantic Drama	Titanic	1997	7
7	Black Comedy	Office Space	1999	133
8	Musical	La La Land	2016	83

Table 1: Films similar to Raiders of the Lost Ark

Most films were selected by popularity as measured by box office gross (The-Numbers.com, 2024), critical reviews (Tomatoes, 2024) and/or pop culture influence (e.g. Tomb Raider video game tie-ins). These criteria ensure most films are well represented in LLM training datasets that include Wikipedia, movie scripts, and movie review websites. The least popular film, Office Space, was included to be used as a stress test check against hallucination as described in section 3.5.

3.2 Comparison Methods and Narrative Source

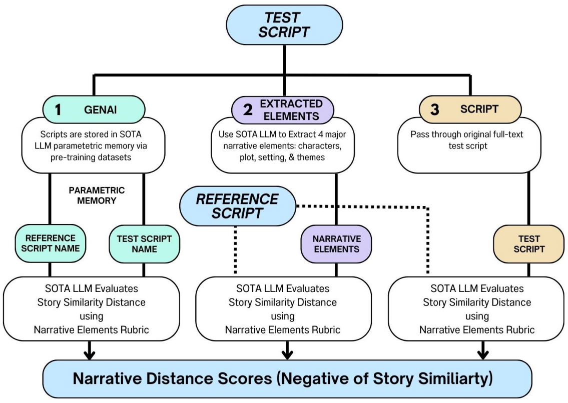

TEST SCRIPT
1 GENAI
Scripts are stored in SOTA LLM parametetric memory via pre-training datasets
2 EXTRACTED ELEMENTS
Use SOTA LLM to Extract 4 major narrative elements: characters, plot, setting, & themes
3 SCRIPT
Pass through original full-text test script
PARAMETRIC MEMORY
REFERENCE SCRIPT
REFERENCE SCRIPT NAME
TEST SCRIPT NAME
NARRATIVE ELEMENTS
TEST SCRIPT
SOTA LLM Evaluates Story Similarity Distance using Narrative Elements Rubric
SOTA LLM Evaluates Story Similarity Distance using Narrative Elements Rubric
SOTA LLM Evaluates Story Similarity Distance using Narrative Elements Rubric
Narrative Distance Scores (Negative of Story Similiarty)

Figure 1: Three Comparison Methods

Each of the 8 test films was compared to the reference film to evaluate the semantic differences using one of three different techniques as shown in Figure 1. First [GenAI]: the SOTA LLM was only provided the names of the reference and test films and asked to evaluate similarity based upon knowledge of both films from parametric memory using the narrative scoring rubric. Second [Elements]: narrative elements and sub-features were extracted from full-text movie scripts and extracts were evaluated for similarity using the narrative scoring rubric. Third[Script]: full-text scripts of both the reference and test films were evaluated for similarity without providing the narrative scoring rubric.

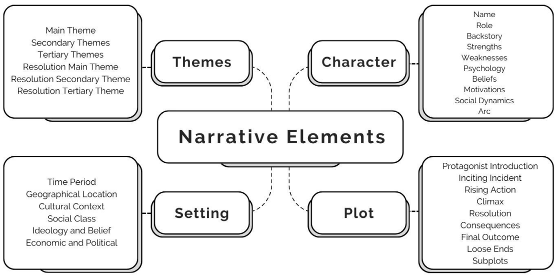


Figure 2: Narrative Similarity Rubric

Similarity comparison methods 1 (Elements) and 2 (GenAI) asked the SOTA LLM-as-judge to provide detailed similarity scores and explanations based on the narrative rubric shown in Figure 2. The extra step to extract and compare individual elements in method 2 Elements was akin to an explicit chain of prompts focusing on a two-step evaluation process. The relative performance of method 1 using only the extracted concise summaries of narrative elements provided an advantage over providing the entire scripts either explicitly (via method 3 Scripts) or implicitly (via method 2 GenAI). Evaluation method 3 (Scripts) was to see how well just providing raw film scripts and relying upon the SOTA LLM to come up with its own similarity evaluation metrics performed. That is, are the SOTA LLMs so capable they need no explicit scoring rubric to perform well?

The four major narrative elements in the scoring rubric consist of 6-10 sub-features as shown in Figure 2. Preliminary tests showed noticeable improvements when decomposing narrative into coherent and focused individual elements over just one large prompt combining all elements and sub-features. The four elements can also be ranked by an approximate order of complexity: Setting (facts), Plot (categorized and properly sequenced events), Character (facts, inferences, and analysis), and Themes (fuzzy categorizations, prioritization, and close readings that require the most abstract thinking and understanding of pragmatics).

The characters narrative element stands out because it contains the most disparate features in terms of type and analysis required. Name, role, backstory, and even strengths/weaknesses are largely factual. Psychology, beliefs and motivations add potentially complex interpretations of characters that are informed not only by descriptions, dialog, and actions but also by constructing mental models of internal personalities and drives that are informed by contextual clues, themes, and more abstract and interrelated sub-features and text. Finally, social dynamics and character arcs add the dimension of time and more interrelated aspects of text and narrative. It's not uncommon for dialog, social dynamics, and arc to be considered separate from characters, but we wanted relatively balanced elements while tracking these character-related topics. Dialog was sufficiently complex and difficult to concisely/comprehensively parameterize as a metric that it was left off in this iteration. Initial tests showed it added significant complexity, prompt task distraction, and resulted in lower signal/noise similarity scoring.

3.3 Models and API

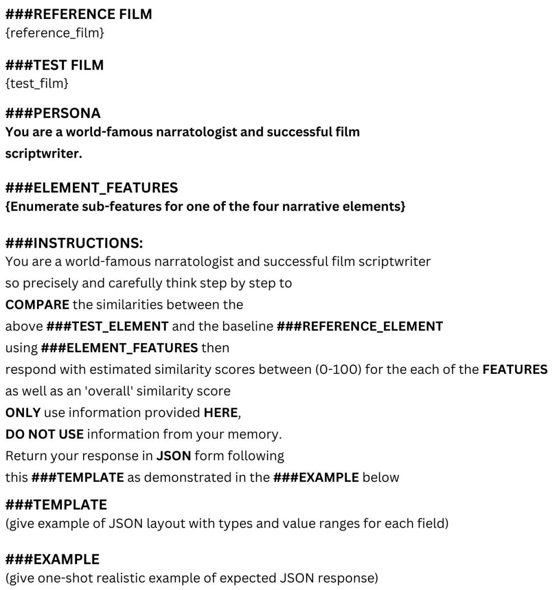
Preliminary testing showed little to no difference between OpenAI gpt-4o and gpt-3.5-turbo, so GPT3.5 was selected as our SOTA LLM used to evaluate similarity for all three scoring methods. It was also used to extract narrative elements in the pre-processing stage for the second comparison method [Elements]. To check against hallucinations, two leading SOTA commercial models at the time of this paper, Claude 3.5 Sonnet and GPT4o, were used to validate factual accuracy as described below. In addition, these two SOTA models were used to provide a naive baseline similarity ranking for all 8 test films with a single prompt (without scripts or a narrative rubric).

Each API call was de novo with no memory or personal history. All OpenAI playground and chat UI interactions had personalization memory disabled and each was submitted afresh after every response to the previous prompt. Prompts were injected with a unique randomized string to avoid possible server-side caching when repeatedly sampling with the same prompt to collect sample sets of $n = 30$. Finally, inference hyperparameters were set as temperature $= 0.7$, top_p $= 0.5$, and response_format = 'json_object'. Initial exploratory analysis of temperature values $= 0.1$, 0.3, and 0.5 did not produce similarity score distributions with informative statistical spread values (e.g. IQR and std) to gauge confidence levels.

3.4 Prompts

Prompts were created to evaluate the semantic similarity between the reference film and 8 test films. The rubric to score overall similarity in Figure 2. is based on 4 main narrative elements and 31 sub-features. Narrative elements include characters, plot, setting and themes with excellent results which each have between 6-10 sub-features as shown. The common anatomy of all prompts is shown in Figure 3 using the 'plot' element. The full text of these four principal prompts can be found in Appendix A.

Two variations of this set of 4 prompts were created: one for evaluation and one for extractions (used only for method 2 Elements). The evaluation prompt asked the LLM to estimate a similarity score (0-100) for each narrative element 'overall' and similarity scores for each of the associated sub-features in Figure 2. LLMs were prompted to provide an open-end 'reason' to justify each similarity score.



###REFERENCE FILM
{reference_film}

###TEST FILM
{test_film}

###PERSONA
You are a world-famous narratologist and successful film scriptwriter.

###ELEMENT_FEATURES
{Enumerate sub-features for one of the four narrative elements}

###INSTRUCTIONS:
You are a world-famous narratologist and successful film scriptwriter
so precisely and carefully think step by step to
COMPARE the similarities between the
above **###TEST_ELEMENT** and the baseline **###REFERENCE_ELEMENT**
using **###ELEMENT_FEATURES** then
respond with estimated similarity scores between (0-100) for the each of the **FEATURES**
as well as an 'overall' similarity score
ONLY use information provided **HERE**,
DO NOT USE information from your memory.
Return your response in **JSON** form following
this **###TEMPLATE** as demonstrated in the **###EXAMPLE** below

###TEMPLATE
(give example of JSON layout with types and value ranges for each field)

###EXAMPLE
(give one-shot realistic example of expected JSON response)

Figure 3: Prompt Template

Eight extractions and comparisons were made to measure the similarity between the reference film and 8 test films. Extractions were run once for all four elements across all 8 reference-test comparisons (32 API calls). Evaluations of story similarity were run 30 times for each 4 narrative elements across all 8 reference-test comparisons for a total of 960 API calls. The cl100k_base tokenizer used by GPT3.5 and GPT4, request token counts varied by comparison method, approximately 1250 for GenAI and 2200 for Elements. Scripts were converted to plain text and attached along with scoring prompts for the Script method.

4 Results

4.1 Overview

The oversized Table 2 in Appendix B compares narrative semantic similarity between the reference film 'Raiders of the Lost Ark' and eight other films. Horizontally, a human expert ordered films left to right from most to least similar in groups of a. two sequels (light yellow), b. three other adventure genre films (medium yellow), and c. three different genre films (dark yellow). Ordering films within each group is based upon the expert's multiple viewing and intimate familiarity of narrative elements. For example, As a epic disaster film, Titanic shares dramatic elements with the adventure genre. The black comedy shares constant sublimated tension and conflict with adventure films. Finally, the relatively emotional and generally light-hearted nature of song-and-dance musical was judged the least similar. Human similarity is simply the rank ordering of similarity distance between each film and the reference film in the row labeled 'Human Similarity-Title'.

In three groupings vertically, AIStorySimilarity's three similarity methods (Elements, GenAI, Scripts) of AIStorySimilarity are compared 1. Across each row, LLM-as-a-judge similarity scores for each method overall as well as broken out by the four constituent narrative elements (character, plot, setting, themes) are listed.

Individual cells give similarity scores (0-100) between the reference film 'Raiders of the Lost Ark' and the film atop each column. The row indicates which combination of 'Similarity Method' and 'Narrative Element' the score corresponds to using the AIStorySimilarity rubric in Appendix A. The similarity score in each cell is based on the mean of 30 samples. Because the Script method proved to be the least reliable, only one film was analyzed from each of the three groups of films (sequels, adventures, and non-adventures) to verify general alignment with human evaluation.

The colored cells in Table 2 highlight the exact points of major differences between human and LLM-as-a-judge similarity ranking using the AIStorySimiliary rubric. These three types of errors are color-coded as follows:

- Red cells indicate similarity scores below human expert ranking
- Green cells indicate scores above human ranking
- Orange cells count as errors to penalize the excessive use of ties

The row of blue cells reflect the overall similarity scores for Elements characters were 80.00 across all models and n=30 iterations. All other Element values appeared correct and well distributed as did characters similarity scores for GenAI and Scripts. Several prompt variations were used to try to correct this, but no OpenAI API response changed this value. We note this anomaly here for completeness and as a point for future investigations.

Surprisingly, the similarity scores least aligned with the human expert are those produced by first extracting all the elements before doing a comparison using method 1 Elements in Table 2. As seen in the similarity plots, this extraction step removes all contextual script information, which results in less nuanced and more narrowly clustered scores. This narrowing of values, combined with both the inevitable information loss in extraction and the inherent noise in natural language descriptions, results in 37.5% total ranking errors compared to human expert ranking. The gaps between different similarity values are dramatically narrowed using the Elements method extraction. Despite numerous misorderings, the magnitude of score differences are relatively small compared to the other two methods.

In contrast, GenAI method similarity scores across all four narrative elements and 8 test films only had 2/32 or 6.25% total errors in ranking. Using a stricter definition of error to mean any misordering to compensate for the reduced test set of only 3 films, the Script method had an approximately equivalent total ranking error rate is 2/12 or 17%.

Based upon overall results in section 4.1, we remove the Elements method from further consideration and focus on comparing the similarity scores from the remaining two methods: GenAI and Scripts. All eight test films' similarity scores are shown in radar charts for both these methods in Figure 4 and Figure 5 respectively. The spokes represent similarity scores for the four narrative elements with the top vertical spoke represents the overall similarity scores.

Despite the better alignment with human experts for this test case, the GenAI method is not a universal solution for measuring story similarity. Notably, GenAI depends upon stories being evaluated that are well represented in the training dataset and parametric memory. Where this is not true (e.g. de novo generated narratives or recently released films after the training date cutoff), the other two methods are required. The choice between the Elements and the Scripts methods involves a series of trade-offs between stability, control, privacy, cost, performance, local edge applications, and other lesser factors.

High-res vector image files of all plots and figures are directly available in the subdirectory at https://github.com/jon-chun/AIStorySimiliarity/data/.

4.2 Comparing Similarity Scores

Results for GenAI in Figure 4 show a nice gradation in similarity score across the test films with "Raiders of the Lost Ark". The eight films gener-

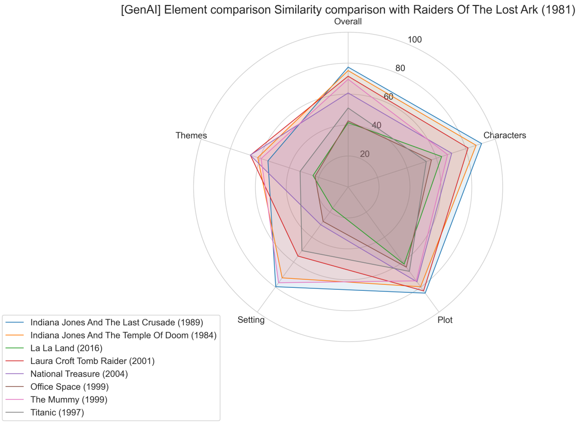


Figure 4: Full GenAI Similarity Scores

ally cluster by similarity in three groupings already noted: two sequels (largest polygons), three unrelated adventure genre films, and the three unrelated genres (smallest polygons). Plot is the most similar narrative element across all films, perhaps due to the near ubiquitous strong hero's journey in Hollywood films targeted at mass audiences. In contrast, Themes reflect the greatest diversity with the lowest similarity scores. This aligns with the earlier idea that Themes is the most abstract, subjective, and artistically unconstrained of the four narrative elements. Most importantly, we get a nice spread along the 'noon' overall similarity axis demonstrating AIStorySimilarity to make both coarse- and fine-grained distinctions between very similar (sequels), similar (adventure), and dissimilar (non-adventure) films.

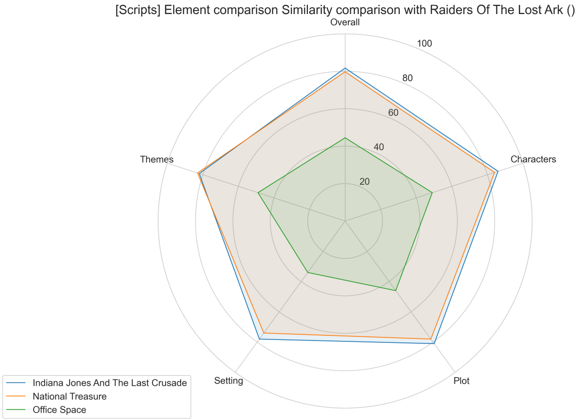


Figure 5: Sampled Script Similarity Scores

Using SOTA LLMs as a judge, the Script method does a relatively good job in similarity scoring when presented with clearly different films as shown in Figure 5. In this case, both the reference and test film scripts were fed into GPT3.5 with no rubric and with only minimal prompting to estimate impromptu similarity scores (0-100). Figure 5 shows a clear distinction between a sequel and adventure film vs a non-adventure film. However, there is poor discrimination between the sequel and adventure film. This suggests that minimalist prompting without an explicit evaluation rubric (e.g. AIStorySimialrity) may be limited to distinguishing between fewer and more distinct films

4.3 Comparing Rankings

The three bar charts in Figure 6 through Figure 8 visualize all 3 methods AIStorySimilarity uses to compute overall similarity scores. As mentioned in section 4.1, the Elements method first decomposes film scripts into distinct narrative elements before scoring. This appears to remove rich contextual information required to draw sharp distinctions. This lowers discrimination power resulting in more ranking errors. In contrast, both generating elements from parametric memory (GenAI) and manually providing copies of scripts (Scripts) result in smoother gradations between films and sharp boundaries between the 3 categories of test films.

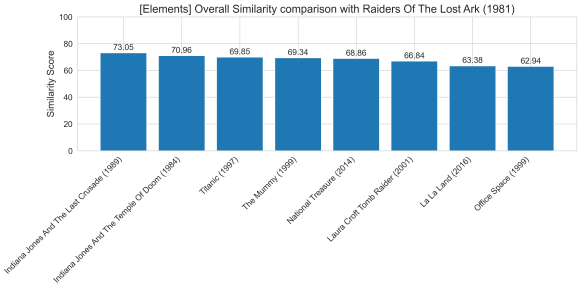


Figure 6: Full Elements Overall Similarity Scores

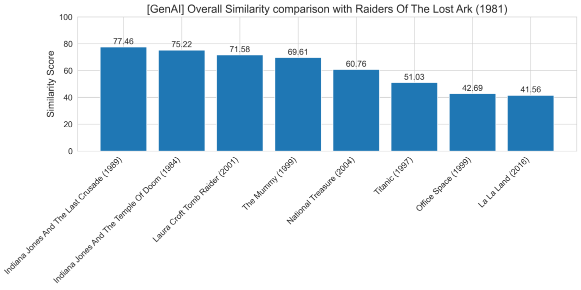


Figure 7: Full GenAI Overall Similarity Scores

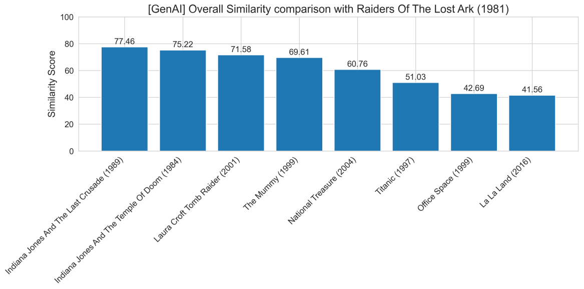


Figure 8: Sampled Scripts Overall Similarity Scores

5 Conclusion

AIStorySimilarity presents a novel story similarity metric and benchmark based upon narratology and best practices in screenplay writing. This benchmark overcomes limitations with traditional text and story similarity metrics and has many potential real-world applications including search/recommendation engines, IP infringement detection, and guided creative AI-collaboration. Three comparisons methods are tested and evaluated including 1. preprocessing scripts to extract concise narrative elements (Elements), 2. using LLM parametric memory with a narrative rubric (GenAI), and 3. providing full-text scripts with a narrative rubric (Scripts). For these famous Hollywood films, the GenAI method proved most aligned with the human expert. However, the other two methods (Elements and Scripts) may be required for narratives that do not exist in parametric memory or are subject to other practical constraints like cost and privacy. In our test dataset, results demonstrate SOTA LLMs have a good innate sense of popular Hollywood films, narrative theory, and can produce results in strong alignment with human experts.

6 Limitations

Three major limitations of this study are the size/-diversity of the film test dataset, the number/size of LLMs tested, and the types of narrative under study. This paper introduced and tested a simplified set of eight test films with clear degrees of similarity to the reference film. With the utility of AIStorySimilarity thus demonstrated, the method should next be stress tested with a much larger and diverse set of test films.

Our current test set did not have enough data or diversity to explore in close detail how our methodology evaluates similarity for semantically very different films or how it distinguishes between a much broader set of genres, or how it categorizes genres and edge cases that are difficult to classify. For example, some genres like musicals and comedies frequently blend aspects of other genres like adventure and romance. Additionally, non-conventional film styles, such as art house, postmodern, and absurdist cinema, are less suited to this approach due to their often fragmented narratives, experimental techniques, and resistance to traditional storytelling conventions.

The strong performance of the commercial SOTA models (GPT3.5, GPT4o and Claude 3.5 Sonnet), raises questions how well small open LLMs can perform under the demands and complexity of interpreting more abstract narrative elements and structures. Finally, measuring the narrative distance for different forms of narratives like those in medical histories, and financial reporting will require customizing the scoring rubric.

This paper limited itself to a focused study of prototypical Hollywood big-budget films across several genres based upon textual scripts. The author is currently expanding this work to work with stories that are multimodal (e.g. video/image, music, and voice) as well as from different cultures and semantic representations.

References

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *Preprint*, arXiv:2401.17072.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Preprint*, arXiv:2406.18403.

Aman Berhe, Camille Guinaudeau, and Claude Barras. 2022. Survey on narrative structure: from linguistic theories to automatic extraction approaches. In *ICON*.

Xavier Bost and Vincent Labatut. 2019. Extraction and analysis of fictional character networks. *ACM Computing Surveys (CSUR)*, 52:1–40.

Brian Boyd. 2017. The evolution of stories: from mimesis to language, from fact to fiction. *Wiley Interdisciplinary Reviews: Cognitive Science*, 9. N. pag.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Ricardo Campos, Alípio Mário Jorge, Adam Jatowt, Sumit Kaur Bhatia, Marina Litvak, João Paulo Cordeiro, Conceição Rocha, Hugo Sousa, and Behrooz Mansouri. 2023. Report on the 6th international workshop on narrative extraction from texts (text2story 2023) at ecir 2023. *ACM SIGIR Forum*, 57:1–12.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have i heard this story before? identifying narrative similarity in movie remakes. In *North American Chapter of the Association for Computational Linguistics*.

Xi Chen, Ali Zeynali, Chico Q. Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw A. Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. Semeval-2022 task 8: Multilingual news article similarity. In *International Workshop on Semantic Evaluation*.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Preprint*, arXiv:2405.13769.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Annual Meeting of the Association for Computational Linguistics*.

Huang Chieh-Yang, Sanjana Gautam, Shannon McClellan Brooks, Ya-Fang Lin, and Ting-Hao 'Kenneth' Huang. 2023. Inspo: Writing stories with a flock of ais and humans. *ArXiv*, abs/2311.16521. N. pag.

Alon Coret, Kerry Boyd, Kevin Hobbs, Joyce Zazulak, and Meghan M. McConnell. 2018. Patient narratives as a teaching tool: A pilot study of first-year medical students and patient educators affected by intellectual/developmental disabilities. *Teaching and Learning in Medicine*, 30:317–327.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. *Preprint*, arXiv:2405.18638.

EU (Parliament) - and Jaume Duch Guillot. 2023. Eu ai act: first regulation on artificial intelligence.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *Preprint*, arXiv:2402.01383.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Annual Meeting of the Association for Computational Linguistics*.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *Preprint*, arXiv:2310.19736.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *Findings*.

Robert F. Helfing. 2020. Substantial similarity and junk science: Reconstructing the test of copyright infringement. *Fordham Intellectual Property, Media & Entertainment Law Journal*, 30:735.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria? *Preprint*, arXiv:2402.12055.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *Preprint*, arXiv:2211.05030.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex 'Sandy' Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024a. Leveraging large language models for learning complex legal concepts through storytelling. *Preprint*, arXiv:2402.17019.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024b. A survey on large language model hallucination via a creativity perspective. *Preprint*, arXiv:2402.06647.

Mayank Jobanputra and Lorena Martín Rodríguez. 2022. Chen et al., 2022. In *International Workshop on Semantic Evaluation*.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. *Preprint*, arXiv:2311.18702.

Christina Lehnen. 2016. Exploring narratives' powers of emotional persuasion through character involvement: A working heuristic. *Journal of Literary Theory*, 10:247–270.

Effi Levi, Guy Mor, Tamir Sheafer, and Shaul Shenhav. 2022. Detecting narrative elements in informational text. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for nlg evaluation: Advances and challenges. *Preprint*, arXiv:2401.07103.

Robert Mckee. 1997. *Story: Substance, Structure, Style*. Reganbooks, New York.

Piotr Wojciech Mirowski, Kory Wallace Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. N. pag.

OpenAI. 2022. Chatgpt. https://chatgpt.com/. Accessed 30 Oct. 2022.

Audrey Pope. 2024. Nyt v. openai: The times's about-face. https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/. Accessed 15 June 2024.

Caroline Rees. 2020. Transforming how business impacts people: Unlocking the collective power of five distinct narratives. *Corporate Governance: Social Responsibility & Social Impact eJournal*. N. pag.

Constanze Schreiner, Markus Appel, Maj-Britt Isberner, and Tobias Richter. 2017. Argument strength and the persuasiveness of stories. *Discourse Processes*, 55:371–386.

Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024. Heart-felt narratives: Tracing empathy and narrative style in personal stories with llms. *Preprint*, arXiv:2405.17633.

Blake Snyder. 2005. *Save the Cat! : The Last Book on Screenwriting You'll Ever Need*. Michael Wiese Productions, Studio City, CA. Accessed 25 May 2005.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So. 2024. Confabulation: The surprising value of large language model hallucinations. *Preprint*, arXiv:2406.04175.

Derek R. Sweet and Margret McCue-Enser. 2010. Constituting "the people" as rhetorical interruption: Barack obama and the unfinished hopes of an imperfect people. *Communication Studies*, 61:602–622.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Preprint*, arXiv:2406.12624.

The-Numbers.com. 2024. Top-grossing movies of 1981. https://www.the-numbers.com/market/1981/top-grossing-movies. Accessed 2 July 2024.

Katrin Tomanek, Jimmy Tobin, Subhashini Venugopalan, Richard Cave, Katie Seaver, Jordan R. Green, and Rus Heywood. 2024. Large language models as a proxy for human evaluation in assessing the comprehensibility of disordered speech transcription. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. N. pag.

Rotten Tomatoes. 2024. Rotten tomatoes: Movies | tv shows | movie trailers | reviews. https://www.rottentomatoes.com/. Accessed 2 July 2024.

John Truby. 2007. *The Anatomy of Story: 22 Steps to Becoming a Master Storyteller*. Farrar, Straus And Giroux, New York.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

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. *Preprint*, arXiv:2303.04048.

Yuxin Wang, Jieru Lin, Zhiwei Yu, Wei Hu, and Börje F. Karlsson. 2022. Open-world story generation with structured knowledge enhancement: A comprehensive survey. *Neurocomputing*, 559:126792.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *International Conference on Natural Language Generation*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *Preprint*, arXiv:2310.01469.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *Preprint*, arXiv:2309.06794.

Dominyk Zdanovic, Tanja Julie Lembcke, and Toine Bogers. 2022. The influence of data storytelling on the ability to recall information. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56:1–37.

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. *Preprint*, arXiv:2306.05685.

Xiaoshi Zhong and Erik Cambria. 2023. Time expression recognition and normalization: a survey. *Artificial Intelligence Review*, pages 1–26.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are nlp models good at tracing thoughts: An overview of narrative understanding. In *Conference on Empirical Methods in Natural Language Processing*.

A Appendix A: Prompt to Compare Narrative Element of Characters

```
###REFERENCE_ELEMENT
{reference_element}

####TEST_ELEMENT:
{test_element}

###PERSONA:
You are a world-famous narratologist and successful film scriptwriter
    .

###ELEMENT_FEATURES
Name: Full name of character
Role: Clarifies the character's function within the story, whether
    they are driving the action, supporting the protagonist, or
    creating obstacles.
Backstory: This attribute helps to understand the formative
    experiences that shaped each character, providing insights into
    their motivations and behaviors.
Strengths: Highlights unique abilities and proficiencies,
    distinguishing characters by their specific talents and expertise.
Weaknesses: Humanizes characters by revealing vulnerabilities and
    personal challenges, making them more relatable and multi-
    dimensional.
Psychology: Uses personality assessments, such as the Big 5 OCEAN (
    Openness, Conscientiousness, Extroversion, Agreeableness,
    Neuroticism) model, to offer deeper insight into character traits.
Beliefs: Offers a window into the ethical and moral framework guiding
     each character's decisions, crucial for understanding their
    actions in moral dilemmas.
Motivations: Describes what drives the character to act, including
    desires, fears, and goals.
SocialDynamics: Explores the nature of interactions between
    characters, which can be pivotal in character development and plot
     progression.
Arc: Summarizes how the character changes or grows for better or
    worse over the story in response to events, decisions, and actions
     taken

###INSTRUCTIONS:
You are a world-famous narratologist and successful film scriptwriter
so precisely and carefully think step by step to
COMPARE the similarities between the attached ###TEST_ELEMENT and the
     baseline ###REFERENCE_ELEMENT
using ###ELEMENT_FEATURES then
responds with estimated similarity scores between (0-100)  for the
    similarity of each of the FEATURES
as well as an 'overall' similarity score
ONLY use information provided HERE,
DO NOT USE information from your memory.
```

```
Return your response in JSON form following this ###TEMPLATE as demonstrated in the ###EXAMPLE below

###TEMPLATE

{
    "overall": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "backstory": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "strengths": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "weakness": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "psychology": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "beliefs": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "motivations": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "social_dynamics": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    },
    "arc": {
        "similarity": integer range(0,100),
        "reasoning": string len(100,200)
    }
}

###EXAMPLE:

{
    "role": {
        "similarity": 90,
        "reasoning": "Both are protagonists who drive the action in pursuit of historical treasures. They lead quests and face adversities while seeking valuable artifacts. The main
```

```
            difference is that Indiana Jones has a more established
            background as an archaeologist and professor."
    },
    "backstory": {
        "similarity": 75,
        "reasoning": "Both characters have backgrounds tied to
            historical pursuits. However, Indiana Jones' backstory is
            more focused on personal experiences shaping his ethical
            stance, while Gates' is deeply rooted in family legacy and
             tradition."
    },
    "strengths": {
        "similarity": 85,
        "reasoning": "Both characters share intelligence,
            resourcefulness, and deep historical knowledge. Indiana
            Jones has additional combat and survival skills, while
            Gates' strengths are more academically focused."
    },
    "weaknesses": {
        "similarity": 70,
        "reasoning": "Both have weaknesses that can lead to reckless
            behavior. Indiana's impulsiveness and fear of snakes are
            more specific, while Gates' obsession with treasure is
            more directly tied to his motivations."
    },
    "psychology": {
        "similarity": 85,
        "reasoning": "They share high openness, conscientiousness,
            and relatively low neuroticism. The main differences are
            in extroversion (Indiana higher) and agreeableness (Gates
            higher)."
    },
    "beliefs": {
        "similarity": 90,
        "reasoning": "Both strongly value history, preservation, and
            protecting artifacts from exploitation. Gates has an
            additional emphasis on familial duty."
    },
    "motivations": {
        "similarity": 80,
        "reasoning": "Both are driven by a desire to preserve history
             and fulfill personal quests. Gates' motivation is more
            focused on family legacy, while Indiana's includes a
            thirst for adventure and living up to his father's legacy
            ."
    },
    "social_dynamics": {
        "similarity": 75,
        "reasoning": "Both form alliances and face adversaries.
            Indiana's relationships are more complex, especially with
            his father and romantic interests. Gates' dynamics focus
            more on his team and main antagonist."
    },
    "arc": {
        "similarity": 85,
        "reasoning": "Both characters evolve to understand deeper
            values beyond their initial quests. Indiana's arc focuses
            on his relationship with his father, while Gates'
            emphasizes valuing relationships and heritage more broadly
            ."
    }
}
```

B Appendix B: Complete Similarity Results

Human Similarity-Title		1-Temple of Doom	2-Last Crusade	3-Tomb Raider	4-The Mummy	5-National Treasure	6-Titanic	7-Office Space	8-La La Land
Similarity Method	Narrative Element	1984 Sequel	1989 Sequel	Adventure	Adventure	Adventure	Drama-Romance	Black Comedy	Musical
Elements	Overall	70.96 (2)	73.05 (1)	66.84 (6)	69.34 (4)	68.86 (5)	69.85 (3)	62.94 (8)	63.38 (7)
	Characters	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00
	Plot	72.50 (4)	71.62 (5)	70.00 (7)	77.06 (1)	76.62 (2)	70.88 (6)	70.00 (8)	73.53 (3)
	Setting	58.82 (3)	70.00 (1)	40.00 (tie 5-8)	60.29 (2)	40.00 (tie 5-8)	57.65 (4)	40.00 (tie 5-8)	40.00 (tie 5-8)
	Themes	72.50 (3)	70.59 (5)	77.35 (2)	60.00 (tie 6-8)	78.82 (1)	70.88 (4)	61.76 (tie 6-8)	60.00 (tie 6-8)
GenAI	Overall	75.22 (2)	77.46 (1)	71.58 (3)	69.61 (4)	60.76 (5)	51.03 (6)	42.69 (7)	41.56 (8)
	Characters	87.06 (2)	90.79 (1)	81.58 (3)	67.67 (5)	70.45 (4)	53.15 (8)	56.67 (7)	63.61 (6)
	Plot	79.73 (3)	84.82 (1)	83.03 (2)	75.03 (5)	75.82 (4)	67.42 (6)	64.06 (7)	61.64 (8)
	Setting	72.64 (3)	79.70 (1)	55.21 (4)	76.52 (2)	30.18 (6)	50.88 (5)	27.55 (7)	17.09 (8)
	Themes	61.45 (3)	54.55 (5)	66.48 (2)	59.21 (4)	66.58 (1)	32.67 (6)	22.48 (8)	23.91 (7)
Scripts	Overall		81.75 (1)			79.75 (2)		44.50 (3)	
	Characters		86.00 (1)			84.00 (2)		49.00 (3)	
	Plot		81.00 (1)			78.00 (2)		46.00 (3)	
	Setting		78.00 (1)			74.00 (2)		34.00 (3)	
	Themes		82.00 (2)			83.00 (1)		49.00 (3)	

Table 2: AIStorySimilarity Scores for Narrative Similarity to 'Raiders of the Lost Ark (1981)'

C Appendix C: Script Dataset Statistics

Film Name	**Characters**	**Words**	**Sentences**	**Vocabulary Size**	**Reading Level**
Raiders of the Lost Ark (1981)	160,278	29,870	2,847	4,730	104
Indiana Jones and the Temple of Doom (1984)	190,111	34,230	2,926	5,142	103
Indiana Jones and the Last Crusade (1989)	137,750	26,181	2,957	4,523	112
Titanic (1997)	246,677	46,028	4,564	6,824	112
The Mummy (1999)	157,912	27,759	3,127	4,571	110
Office Space (1999)	64,777	12,838	1,661	2,037	118
Lara Croft Tomb Raider (2001)	158,941	28,546	2,479	5,678	106
National Treasure (2004)	169,878	31,030	3,485	5,113	119
La-La-Land (2016)	104,568	20,520	2,416	3,626	114

Table 3: Simplified Scripts Dataset Statistics