Entertainment Computing 58 (2026) 101166

Contents lists available at ScienceDirect

Entertainment Computing

journal homepage: www.elsevier.com/locate/entcom

# Trait-guided detective story generation in seven classic investigative styles with Large Language Models ☆

Edirlei Soares de Lima [a,*], Margot M.E. Neggers [a], Marco A. Casanova [b], Bruno Feijó [b], Antonio L. Furtado [b]

[a] Breda University of Applied Sciences, Academy for AI, Games and Media, Monseigneur Hopmansstraat 2, Breda, 4817 JS, The Netherlands
[b] Pontifical Catholic University of Rio de Janeiro, Department of Informatics, Rua Marquês de São Vicente 225, Rio de Janeiro, 22451-900, Brazil

ARTICLE INFO

*Keywords:*
Detective fiction
Narrative generation
Large language models
Interactive storytelling

ABSTRACT

Detective stories are a particularly popular form of mystery fiction driven by investigative reasoning and the gradual revelation of hidden truths. This work presents a trait-guided method for generating detective stories that reflect the investigative styles of classic fictional detectives while maintaining narrative quality. Building upon a preliminary study that characterized the investigative methods of seven iconic detectives (Hercule Poirot, Sherlock Holmes, William Murdoch, Columbo, Father Brown, Miss Marple, and Auguste Dupin), we developed a new narrative generation approach that uses formalized investigative trait profiles and Large Language Models (LLMs) to write detective stories. The method integrates these profiles with narrative principles grounded in literary theory, particularly Todorov's dual narrative structure, to produce stories where selected detectives solve crimes according to their characteristic investigative approaches. We conducted a two-phase evaluation study with 210 trait-guided stories to assess both the recognizability of detective-specific investigative styles and the overall narrative quality of the generated stories. In Phase 1, state-of-the-art language models achieved 82.4% accuracy in identifying detectives from anonymized narratives based on investigative characteristics, with 78.8% accuracy when restricted to investigative methods alone, demonstrating that the generated stories successfully embedded recognizable investigative styles. In Phase 2, LLM-based quality evaluation revealed that trait-guided stories achieved higher scores than stories generated without trait guidance on empathy, coherence, surprise, and complexity, with length-controlled analyses confirming robust condition effects on empathy and coherence, and particularly strong improvements for psychologically-oriented detectives. A preliminary human validation supported the central empathy finding and indicated that the trait-guided method's benefits on surprise and engagement may be more pronounced than the LLM scores alone suggest. These findings demonstrate that computational methods can generate detective narratives that preserve distinctive investigative styles while enhancing character depth and emotional portrayal, contributing to the field of computational narratology and AI-assisted creative writing.

## 1. Introduction

Detective stories are a particularly popular type of mystery fiction, which Marie-Laure Ryan characterizes as the standard representative of *epistemic narratives*, that is plots "driven by the desire to know" [1]. These narratives engage the reader not merely by recounting events but by challenging them to solve a puzzle: distinguishing true and false clues, and uncovering hidden truths before they are revealed by the narrative itself. Ryan contrasts this relatively modern genre with two classical forms explored by Aristotle [2]: *epic narratives*, centered on physical heroic action, and *dramatic narratives*, focused on interpersonal human relationships.

The epistemic structure of detective fiction often revolves around the investigator's ability to interpret subtle traces left behind by a crime. This motif finds its early roots in the Persian tale *The Three Princes of Serendip*, in which the protagonists demonstrate remarkable sagacity by making discoveries from seemingly trivial observations. Inspired by this story, Horace Walpole coined the term *serendipity* in 1754 to describe "discoveries, by accidents and sagacity, of things which they were not in quest of" [3]. This allusion to accidental discovery highlights a fundamental tension in detective work: clues are necessary but insufficient on their own, as they only become meaningful through the investigator's capacity to interpret them.

☆ This article is part of a Special issue entitled: 'ENTCOM_ICEC 2026 Journal Track papers' published in Entertainment Computing.

* Corresponding author.
*E-mail address:* soaresdelima.e@buas.nl (E.S. de Lima).

https://doi.org/10.1016/j.entcom.2026.101166
Received 14 January 2026; Received in revised form 7 May 2026; Accepted 10 June 2026
Available online 30 June 2026

A defining aspect of any detective work is the form of reasoning used to interpret clues. Investigators must ask what past events could plausibly explain the observed effects. This approach mirrors medical diagnosis, where symptoms are linked to potential conditions through the formulation and elimination of hypotheses. Although reasoning from effects to causes does not follow the strict rules of deductive logic, this inferential process is formally recognized as *abduction*, a concept introduced by the semiotician Charles Sanders Peirce [4]. Abductive reasoning is widely regarded as central to detective fiction, capturing the logic behind what Edgar Allan Poe described as "retrograde operations" in the case of Auguste Dupin and what Arthur Conan Doyle framed through Sherlock Holmes as "reasoning backwards" [5].

Another characteristic of detective stories is their dual narrative structure. As noted by the literary scholar Tzvetan Todorov [6], these narratives consist of two distinct yet interrelated stories: the *story of the crime*, which describes past events, and the *story of the investigation*, which unfolds in the present and gradually uncovers the truth through logical reasoning and deduction. While the story of the investigation emphasizes intellectual problem solving, detective stories often incorporate episodes of physical action and human relationship dynamics, which are elements that, as noted by Ryan [1], predominate respectively in epic and dramatic narratives. These elements appear in varying degrees across epistemic narratives, enriching their structure and reflecting the diverse investigative approaches portrayed in the genre.

Computational narratology has demonstrated that formalizing narrative structures and character-specific traits can meaningfully guide automated story generation across a range of domains. Structured approaches have been applied to enforce genre conventions in murder mystery scenarios [7,8], to model narrative suspense [9,10], and to improve coherence and character consistency in LLM-based story generation [11,12]. Despite these advances, the systematic characterization of the distinctive investigative methods employed by fictional detectives and their use as structured constraints for narrative generation remains largely unexplored. Motivated by this gap, we previously developed a multi-phase process to formalize the investigative approaches of fictional detectives in a format suitable for narrative generation [13]. As a result, we extracted and validated detailed trait profiles for seven iconic investigators: Hercule Poirot, Sherlock Holmes, William Murdoch, Columbo, Father Brown, Miss Marple, and Auguste Dupin. These profiles captured each detective's distinctive reasoning strategies and behavioral patterns, providing a foundation for generating stories that authentically reflect their investigative styles.

In this work, we take a step further by using the identified investigative traits to generate detective stories that reflect each character's distinctive style. The proposed method takes as input a user-defined crime premise and a selected fictional detective, and produces a narrative in which the chosen investigator solves the case. The generation process is guided by the detective's trait profile, which informs how clues are discovered, how suspects are interrogated, and how conclusions are reached. A Large Language Model (LLM) is used to compose the story, ensuring that the detective's reasoning and behavior remain consistent with their established narrative identity. To validate this approach, we conducted a two-phase evaluation study. Phase 1 assessed whether the generated stories contain recognizable detective-specific investigative characteristics through a multi-LLM identification task. Phase 2 evaluated narrative quality by comparing trait-guided stories against baseline stories generated without investigative trait guidance across six established narrative quality criteria.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 presents an overview of the LLM-based process used in our preliminary study to characterize the investigative styles of fictional detectives. Section 4 introduces the trait-guided detective story generation method and presents the implementation of a functional prototype for interactive story composition. Section 5 presents the two-phase evaluation study examining investigative style recognizability and narrative quality. Section 6 contains concluding remarks.

## 2. Related work

Detective story generation has been explored through a range of structured and procedural methods. Early systems often relied on logic-based or rule-driven approaches to enforce genre conventions. Barbosa et al. [7] use logic programming in Prolog to formally encode genre rules, enabling plan-based plot composition with interactive authoring tools. Jaschek et al. [8] apply linear logic and Monte Carlo Tree Search to simulate believable agent behaviors in murder mystery scenarios. Mohr et al. [14] explore the use of dynamic epistemic logic to track player knowledge, ensuring the solvability of procedurally generated murder cases. Barros et al. [15] construct murder mysteries using open data mined from Wikipedia, relying on evolutionary search to ensure narrative diversity and solvability. A different approach is explored in Cheong and Young's Suspenser system [9], which manipulates reader uncertainty by modeling comprehension through plan-based reasoning. Similarly, Doust and Piwek [10] propose a formal model of suspense using narrative threads and imminence, distinguishing between conflict-based and revelatory suspense. While these approaches establish structural foundations for detective story generation, they rely primarily on explicit logical constraints and generally lack mechanisms for representing the distinctive investigative styles and reasoning patterns that characterize individual fictional detectives.

More recent studies on detective stories have examined the application of LLMs for narrative generation and analysis. Zhao et al. [16] introduced a benchmark to evaluate the ability of LLMs to infer character relationships within detective stories, finding that LLMs struggle with tracking complex interpersonal dynamics and temporal reasoning. These results highlight the difficulty LLMs face in relational reasoning and underscore the need for structured narrative theories to guide generation. Xu et al. [17] present DetectiveQA, a benchmark for long-context reasoning in detective novels, which requires models to track clues and suspect information across narratives averaging over 100,000 tokens. Their evaluation of mainstream LLMs demonstrates persistent challenges in evidence retrieval and reasoning across extended narratives, particularly in maintaining coherence between early clues and final revelations. Wagner et al. [18] propose a probabilistic framework for detective fiction that formalizes the tension between narrative coherence and surprise, introducing the concept of fair play as a quality metric. Using multiple LLMs as both generators and evaluators, their analysis reveals that while LLM-generated stories may be unpredictable, they generally fail to balance surprise with coherent clue structures, suggesting fundamental limitations in current generation approaches. Xie and Riedl [19] propose an iterative prompting technique for suspenseful story generation using LLMs, drawing on cognitive and narratological theories to incrementally construct plotlines with increasing tension. Their approach demonstrates how theoretical frameworks can guide generation, though it focuses on suspense mechanics rather than character-specific styles. Similarly, Iudin [20] combines centralized narrative planning with emergent agent behavior, incorporating LLMs to generate natural language expressions for procedurally created detective stories. While these works advance detective story generation through improved prompting techniques, evaluation benchmarks, and quality metrics, none of them incorporate the distinctive investigative methods of specific fictional detectives as guiding constraints for generation. Our work complements these approaches by using the investigative traits of fictional detectives as structured templates that inform both the reasoning process and the narrative progression in LLM-generated stories.

Beyond the detective domain, broader research in LLM-based story generation has explored methods to improve narrative coherence and structure. Approaches focused on character consistency and narrative control include fine-tuning on large-scale datasets [21,22], reinforcement learning for goal-directed storytelling [23], and knowledge retrieval to support logical consistency [24]. Techniques for refining generation include backward reasoning [25], recursive reprompting [26],

alignment with authorial intent [11], and beat structuring for interactive storytelling [12]. Recent advances include reasoning-based approaches that apply reinforcement learning to plan chapter-level narrative structure for long-form generation [27], and plot-planning frameworks that emphasize coherence through structured event representation, using subject-verb-object triplets to track plot progression [28]. Multi-LLM collaborative systems have explored how multiple LLMs can co-author stories and how to identify each model's contributions [29], though these focus on collaborative writing processes rather than embedding character-specific investigative styles as narrative constraints. Complementary work on narrative pattern reuse has demonstrated that LLMs and computational systems can effectively use archetypal plot structures and story templates [30–33]. Additionally, semiotic approaches to narrative composition have explored how new stories can be systematically derived from existing narratives through operations such as combination, imitation, reversal, or expansion along structural axes [34–36], offering potential mechanisms for story transformation beyond character-driven generation. While these methods advance the technical capabilities of story generation systems, they primarily focus on general narrative structures or underlying model capabilities rather than character-centric frameworks. Our approach, by contrast, applies investigative traits drawn from fictional detectives as structured constraints that inform how clues are discovered, how suspects are interrogated, and how conclusions are reached, ensuring that each detective's reasoning and behavior remain consistent with their established narrative identity.

## 3. The investigative methods of fictional detectives

In a preliminary study [13], we aimed to characterize the investigative methods employed by fictional detectives using a multi-LLM approach. Our method was designed to extract, synthesize, and validate the distinctive investigative traits that define each detective's approach to solving crimes. Fundamentally, this approach seeks to "investigate the investigators", exploring the generative and analytical capabilities of multiple LLMs to identify the essential components of each investigative method.

A sample of seven fictional detectives was selected for this study. The selection was based primarily on the detectives' cultural significance and widespread recognition, as evidenced by their prominent roles in widely read novels and/or successful television adaptations. Furthermore, the selection aimed to offer a balanced representation of the major types of fictional investigators, reflecting the diversity of approaches found within the genre:

- Private investigators: Sherlock Holmes (Arthur Conan Doyle), Hercule Poirot (Agatha Christie).
- Amateur detectives: Miss Marple (Agatha Christie), Father Brown (G. K. Chesterton), Auguste Dupin (Edgar Allan Poe).
- Police detectives: Columbo (Richard L. W. Link), William Murdoch (Maureen Jennings).

The process to characterize the investigative methods is structured around an LLM-based workflow comprising five sequential phases: (1) Description Generation, in which each LLM independently generates concise textual descriptions capturing distinctive investigative traits for each detective; (2) Trait Extraction, where each description is processed by an LLM to extract a structured list of investigative method components; (3) Semantic Grouping, where these components are clustered based on their semantic similarity; (4) Consistency Analysis and Trait Synthesis, which quantitatively evaluates consensus among models, retaining only the most consistently identified traits representative of each detective's investigative methods; and (5) Validation via Reverse Identification, which verifies the distinctiveness of the synthesized traits by requiring LLMs to identify detectives solely from their trait descriptions.

The multi-LLM approach was adopted to enhance the reliability and generalizability of the results. By drawing on the outputs of fifteen LLMs, each with potentially distinct training corpora and reasoning behaviors, we aimed to diversify interpretive perspectives and reduce potential biases or hallucinations from individual models. The ensemble of LLMs functioned as a distributed evaluative framework, increasing the likelihood that consistently emerging traits reflect broadly recognized patterns in the investigative methods of fictional detectives. This strategy also allowed us to explore a form of automated literary analysis, grounded in the collective knowledge of detective fiction encoded across different LLMs.

Our practical experiment with the proposed approach confirmed that the initial three phases of the workflow indeed generated, as expected, structured sets of candidate traits that described each detective's investigative method. These traits were grouped by semantic similarity and served as input for the fourth phase, which applied a quantitative consistency threshold to identify the most reliably extracted traits. The resulting profiles, composed only of traits supported by multiple independent models, establish the foundation for the story generation method presented in this work. In the fifth phase, we assessed the distinctiveness of each profile through reverse identification using fifteen LLMs, achieving an overall accuracy of 91.43%. Several detectives, including Hercule Poirot, Columbo, Father Brown, and Miss Marple, were identified with perfect accuracy, demonstrating the clarity and uniqueness of their synthesized profiles. Cases of misclassification were rare and generally explained by overlapping features between detectives. As a final validation step, we compared the synthesized traits against established literary analyses, confirming their alignment with theoretical interpretations. A complete analysis of the results is presented in our preliminary open-access study [13]. The prompts used across the phases of this workflow are provided in Appendix A for completeness.

The full set of investigative methods used as the foundation for the story generation approach described in this paper is presented in Appendix B.

## 4. AI-assisted detective story composition

Building on the investigative traits derived from our earlier analysis, we developed a story generation method designed to produce coherent narratives that reflect the style of a selected fictional detective. The method applies the synthesized trait profiles as guiding constraints for narrative composition. Rather than relying on fine-tuning or multi-stage pipelines, our approach uses a prompt-based mechanism to guide the narrative generation process through detective-specific investigative traits. This allows us to evaluate the generative capabilities of general-purpose LLMs in the domain of detective fiction, while maintaining transparency over the instructions and constraints guiding the LLM. The following subsections describe the conceptual design of this trait-guided method and its implementation in a fully functional prototype.

### 4.1. Investigative trait-guided story generation

The story generation process proposed in this work is powered by an LLM and guided by a parameterized prompt that integrates two key inputs: the name of a fictional detective (Hercule Poirot, Sherlock Holmes, William Murdoch, Columbo, Father Brown, Miss Marple, or Auguste Dupin), and a textual premise outlining the central crime to be investigated. These parameters are used to dynamically build a story generation prompt, which is designed to structure the narrative in accordance with the selected detective's investigative methods. By grounding the generation process in these predefined traits, the method enables a comparative exploration of how different fictional detectives might approach the same case, highlighting contrasts in reasoning, tone, and narrative progression. This structure facilitates the creation of stylistically distinct detective stories for the same crime premise.

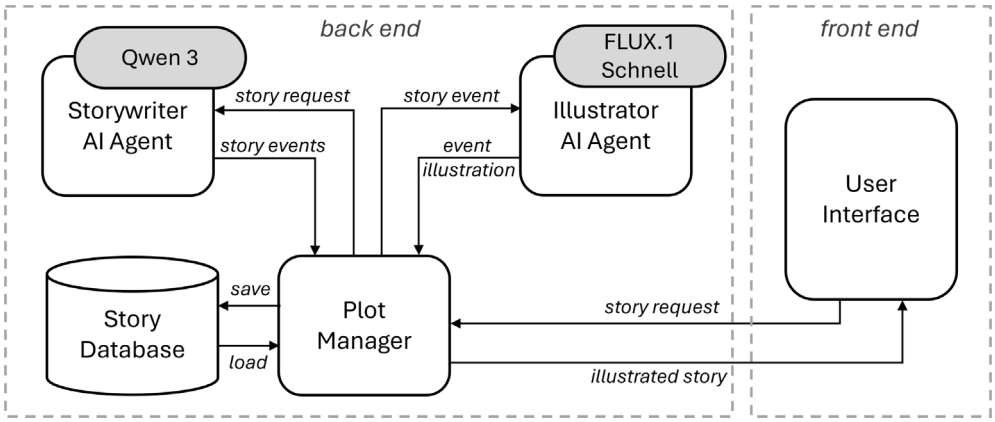


**Fig. 1.** The architecture of the AI-Sleuths prototype.

The story generation prompt, presented in Appendix C, was designed to guide the LLM in generating stories that not only align with the detective's investigative approach but also adhere to narrative principles grounded in literary theory. Central to this design is the operationalization of Todorov's dual narrative structure [6], which encourages the LLM to reveal the hidden story of the crime gradually through the unfolding investigation. This framing reinforces investigative logic while preserving narrative suspense. Additional elements in the prompt specify the presence of clues, inconsistencies, red herrings, and abductive reasoning, all of which are key ingredients in classical detective fiction [5,37]. Dialogue is emphasized as a means of revealing character motivation and conducting interrogation, consistent with conventions of the genre where interaction often drives revelation and misdirection [38].

In addition to narrative guidance, the prompt specifies a consistent output format to ensure that the generated stories can be parsed for a structured visual presentation. Each story begins with a title, followed by a sequence of numbered narrative events, each accompanied by a corresponding event description and scene illustration description. This structured format serves both narrative and practical purposes: it enforces a coherent progression of scenes while enabling the automatic identification and extraction of story components. Markdown-style formatting is used to enhance the readability of the narrative events. By constraining the length to approximately 12 to 18 events, the prompt encourages the LLM to generate concise yet complete storylines, suitable for comparison across different detective styles and crime premises. This range was chosen to provide sufficient narrative space for the detective's investigative process to unfold meaningfully while remaining manageable for evaluation across a large dataset of generated stories.

### 4.2. Story refinement via user suggestions

While the trait-guided generation process produces coherent detective stories aligned with each character's investigative methods, users may desire to modify specific narrative events to better suit their creative vision or address particular story elements. To accommodate this requirement, we introduced a refinement method that allows users to suggest modifications to individual events while preserving the overall narrative coherence and the detective's characteristic investigative approach.

The refinement process operates on the structured format established in the initial generation phase. Any specific event within the generated narrative can be targeted for modification through textual suggestions describing the desired changes, ranging from minor adjustments to character dialogue to more substantial alterations in plot development or investigative discoveries.

When the modification of a particular event is requested, the method employs a specialized regeneration prompt (presented in Appendix D) that takes as input the complete original story, the parameters used for its generation, the target event index, and a description of the suggested modification. The prompt instructs the LLM to regenerate the specified event while maintaining consistency with the detective's investigative traits and preserving logical coherence with all preceding events. When modifications potentially affect subsequent story elements, the later events are adapted to ensure narrative consistency throughout the complete storyline. The prompt also incorporates the narrative guidance principles established in the initial generation phase, maintaining Todorov's dual narrative structure and the established conventions of detective fiction.

### 4.3. The AI-Sleuths prototype

To demonstrate the practical application of our AI-assisted detective story generation method, we developed *AI-Sleuths*, a fully functional prototype designed primarily for creative writers and casual interactive storytelling enthusiasts seeking to explore and experiment with the distinctive investigative styles of classic fictional detectives, as well as for researchers investigating computational approaches to character-driven narrative generation. The system supports creative ideation and stylistic exploration by enabling users to generate, compare, and iteratively refine detective narratives across different investigative styles and crime scenarios.

The architecture of AI-Sleuths follows a multi-agent approach for story composition and visualization, as illustrated in Fig. 1. The system implements two specialized AI Agents: (1) the Storywriter AI Agent, which handles the generation of detective stories according to the trait-guided method described in the previous sections, processing crime premises and investigative styles of different detectives to produce structured narrative sequences that include both story events and corresponding scene descriptions; and (2) the Illustrator AI Agent, which provides visual storytelling support by transforming the textual scene descriptions into illustrative images that depict key moments, characters, and settings for each story event.

Central to the architecture of the prototype is the Plot Manager, which serves as the primary coordination component that manages the narrative generation workflow. The Plot Manager processes user inputs, including crime premises and detective selections, coordinates the story generation process with the Storywriter AI Agent, and manages the subsequent image generation for each story event using the services of the Illustrator AI Agent. All generated detective stories are stored in a dedicated Story Database for subsequent access.

The prototype implements a modular plugin-based architecture for the AI Agents, which enables seamless integration of different LLMs and image generation models. In our current implementation, the

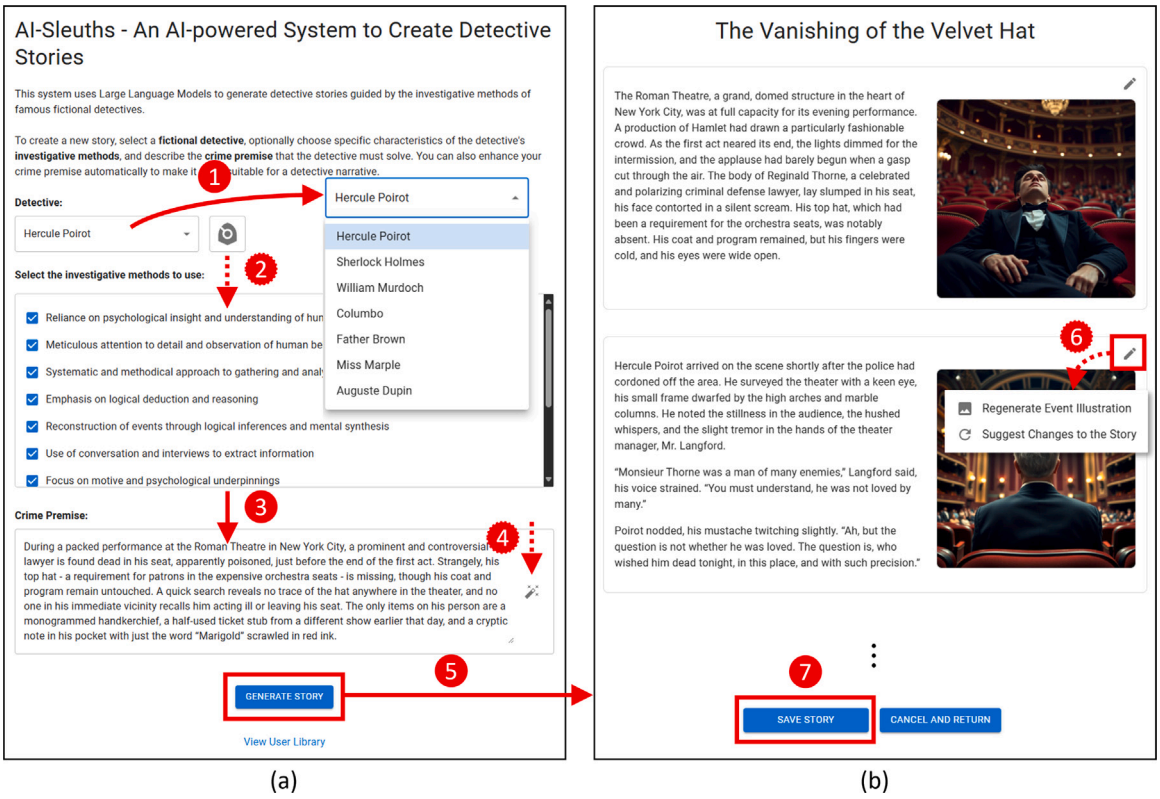


**Fig. 2.** The user interface of AI-Sleuths.

Storywriter AI Agent uses the Qwen 3 32B model for story generation,[1] selected for its strong performance in creative writing tasks and logical reasoning capabilities [39], combined with practical advantages as an open-source model that can be deployed locally on consumer hardware, avoiding the costs and API dependencies associated with proprietary alternatives. Preliminary testing indicated that Qwen 3 32B demonstrated superior instruction-following and narrative generation quality among comparably-sized open models, producing engaging detective stories while maintaining reasonable low computational requirements. The Illustrator AI Agent employs the FLUX.1 Schnell model for image generation,[2] which was chosen for its ability to generate high-quality images with rapid inference times.

To improve the visual consistency across narrative events, the system generates illustrations through a multi-step process. During story generation, the Storywriter AI Agent produces both narrative event descriptions and corresponding image descriptions for each scene. These textual outputs are then processed by the Illustrator AI Agent to create visual representations. The image generation workflow uses the image description to define the visual composition of each scene, while extracting physical characteristics of characters from the event narrative to enrich the prompt with additional details. The system maintains a record of character attributes throughout the story, such as clothing, age group, and hair style, enabling consistent incorporation of these characteristics into the image generation prompts for all subsequent events. To enhance facial consistency and overall character appearance, the system employs name substitution strategies that replace character names with those of recognizable actors familiar to the text-to-image model. Additionally, stylistic parameters are applied to ensure a cohesive visual aesthetic throughout the story, and negative prompts are used to avoid unwanted artifacts. The optimized prompts are then processed by the FLUX.1 Schnell model to generate the final illustrations.

User interaction is performed through a web-based interface that facilitates story composition and narrative visualization (Fig. 2). Users begin the story generation process by selecting one of the seven fictional detectives from the available options (Action 1 in Fig. 2(a)). Optionally, users may customize which specific investigative traits should guide the narrative generation (Action 2 in Fig. 2(a)), allowing for experimentation with different combinations of investigative characteristics. The crime premise is then entered in the designated area (Action 3 in Fig. 2(a)). To assist users in writing detailed and engaging premises, the interface provides an optional "Enhance Crime Premise" feature (Action 4 in Fig. 2(a)), which uses an LLM to enrich the premise description with additional complexity and investigative opportunities while preserving the core elements of the original input. The complete enhancement prompt is presented in Appendix E. Once the detective and premise are defined, clicking the "Generate Story" button (Action 5 in Fig. 2(a)) initiates the narrative generation process, transiting to the story visualization screen.

The story visualization interface, shown in Fig. 2(b), displays the generated detective story as a sequence of illustrated narrative events. Each event presents the textual description alongside its corresponding visual illustration. For each story event, users have the option to regenerate the illustration or suggest modifications to the narrative content (Action 6 in Fig. 2(b)), enabling iterative refinement of the generated story according to the process described in Section 4.2. Upon completing the story, users can save their work by clicking the "Save Story" button (Action 7 in Fig. 2(b)), which stores the narrative in the Story Database for future access.

The AI-Sleuths prototype is accessible through our public website: https://narrativelab.org/ai-sleuths/. An example of story, titled "The Vanishing of the Velvet Hat", whose initial events are presented in Fig. 2(b) is available at: https://narrativelab.org/ai-sleuths/#/story/318.

[1] https://huggingface.co/Qwen/Qwen3-32B.
[2] https://huggingface.co/black-forest-labs/FLUX.1-schnell.

## 5. Evaluation and results

To assess the effectiveness of our trait-guided detective story generation method, we conducted a two-phase evaluation study. The first phase examines whether the generated stories authentically reflect the distinctive investigative styles of the selected fictional detectives. The second phase evaluates the overall narrative quality of the trait-guided stories compared to a simplified prompt baseline, isolating the effect of trait guidance on narrative quality. The following subsections present the methodology and results for each evaluation phase.

### 5.1. Phase 1: Detective representation

In the first phase of this study, we applied a detective identification procedure to determine whether the investigative characteristics embedded in the generated narratives are sufficiently distinctive and recognizable. This evaluation involves generating a set of stories for the seven fictional detectives using identical crime premises, removing detective-specific identifiers from the narratives, and using LLMs to identify which detective's investigative style is represented in each story. Successful identification demonstrates that the proposed method effectively incorporates the characteristic investigative approaches of different fictional detectives into the generated narratives.

#### 5.1.1. Premise formulation

To evaluate the detective identification capabilities across different narrative contexts, three crime premises were formulated for this study. These premises were deliberately designed to represent diverse investigative scenarios, drawing inspiration from classic detective fiction while providing sufficient complexity to accommodate the distinct investigative approaches of different fictional detectives. The following premises were used:

- Premise 1: *"During a packed performance at the Roman Theater in New York City, a prominent and controversial lawyer is found dead in his seat, apparently poisoned, just before the end of the first act. Strangely, his top hat – a requirement for patrons in the expensive orchestra seats – is missing, though his coat and program remain untouched. A quick search reveals no trace of the hat anywhere in the theater, and no one in his immediate vicinity recalls him acting ill or leaving his seat. The only items on his person are a monogrammed handkerchief, a half-used ticket stub from a different show earlier that day, and a cryptic note in his pocket with just the word "Marigold" scrawled in red ink".*
- Premise 2: *"In a dangerous area of Paris, a cry is heard from the Poivrière bar. Inside, two men lie dead near the fireplace, and another is found in the center of the room. A wounded man, clearly the killer, stands in a doorway. He claims innocence, insisting it was self-defense, before attempting to flee. When captured, he exclaims, "Lost... It is the Prussians who are coming". The dying wounded man accuses Jean Lacheneur of luring him to the location and vows revenge. The dead man's attire identifies him as a soldier, with his regiment's name and number marked on his greatcoat's buttons".*
- Premise 3: *"On the morning of the annual harvest festival, the mayor's daughter is found dead at the center of the abandoned village church. Her body is arranged atop the altar and surrounded by ritual symbols. The building had been sealed for decades, with its only door secured by a heavy chain and padlock on the outside, and the snow around it left completely undisturbed. Her last confirmed sighting was at 2:00 a.m. in the town square, where she said goodnight to friends under the observation of both security cameras and the village constable. Just before 3:00 a.m., during a brief power outage, several villagers reported hearing the ancient church bell ring once, despite the bell's mechanism having been declared inoperable many years ago. At dawn, when the door was found still locked, there were no signs of forced entry, no broken windows, and no footprints anywhere near the church. Inside, the only clues are a single scorched corn doll at the girl's feet and a shattered pocket watch, its hands stopped at exactly 3:17 a.m".*

Premise 1 is based on *The Roman Hat Mystery* (1929) by Ellery Queen, which exemplifies the fair-play mystery tradition where the author provides the reader with all the clues necessary to solve the crime, allowing the reader to logically deduce the solution alongside the detective. This premise emphasizes physical evidence, missing objects, and cryptic messages, providing opportunities for both deductive reasoning and psychological insight. Premise 2 draws from *Monsieur Lecoq* (1869) by Émile Gaboriau, considered one of the earliest detective novels and a precursor to modern forensic investigation stories. This premise presents a crime scene with multiple victims, conflicting testimonies, and physical evidence requiring interpretation, allowing for various investigative approaches from forensic analysis to interrogation strategies. Premise 3 is an original formulation based on the locked-room mystery trope,[3] a subgenre popularized by authors such as John Dickson Carr and frequently employed by Agatha Christie. This premise emphasizes seemingly impossible circumstances, temporal inconsistencies, and ritualistic elements, challenging detectives to reconcile physical impossibilities with logical explanation.

#### 5.1.2. Narrative generation

To generate the narratives for this study, we utilized the AI-Sleuths prototype to produce detective stories based on the three selected premises and the investigative styles of the seven fictional detectives: Hercule Poirot, Sherlock Holmes, William Murdoch, Columbo, Father Brown, Miss Marple, and Auguste Dupin. The story generation process was conducted through direct API (Application Programming Interface) requests to the AI-Sleuths backend, ensuring consistent application of the trait-guided generation method across all narratives.

For each detective and premise combination, AI-Sleuths was used to generate 10 independent stories, resulting in a dataset of 210 narratives (7 detectives × 3 premises × 10 stories). The generation process followed the methodology described in Section 4.1, where the Storywriter AI Agent (Qwen 3 32B model) received as input the crime premise, the detective's name, and the detective's complete set of investigative characteristics as identified in our previous study [13]. The generation process was fully automated, with the system producing each story event sequentially according to the structured format specified in the story generation prompt (Appendix C), without any human intervention or modifications during the generation phase. For the purposes of this evaluation, only the narrative event descriptions were utilized, as the image descriptions serve solely as input for the illustration generation process.

All generated stories were preserved in their original form for subsequent analysis. These narratives, which we refer to as the *original stories*, represent the direct output of the trait-guided generation method and serve as the foundation for the detective identification evaluation described in the following subsections.

The complete dataset, including all original stories along with metadata specifying the detective, premise, and investigative characteristics used during generation, is publicly available at: https://narrativelab.org/ai-sleuths-dataset/.

#### 5.1.3. Narrative anonymization

To evaluate whether LLMs can identify detectives based solely on their investigative styles rather than explicit identifiers, all original stories underwent an anonymization process to remove direct references to the detective's name, gender, and associated contextual elements. This process was conducted using OpenAI's GPT-4o model, which was selected after preliminary testing of both open and proprietary LLMs.

[3] https://tvtropes.org/pmwiki/pmwiki.php/Main/LockedRoomMystery.

**Table 1**
Overview of the LLMs evaluated in Phase 1 for detective identification.

| Model | Description | Version |
|---|---|---|
| Claude Opus 4.1 | Flagship model for complex tasks by Anthropic. | claude-opus-4-1-20250805 |
| Claude Sonnet 4 | Mid-tier model by Anthropic. | claude-sonnet-4-20250514 |
| Claude Sonnet 4.5 | Balanced performance model by Anthropic. | claude-sonnet-4-5-20250929 |
| DeepSeek R1 (671B) | 671-billion-parameter reasoning model by DeepSeek. | deepseek-r1:671b |
| DeepSeek R1 (70B) | 70-billion-parameter reasoning model by DeepSeek. | deepseek-r1:70b |
| Gemini 2.0 Flash | Lightweight model by Google. | gemini-2.0-flash |
| Gemini 2.5 Flash | Fast and efficient model by Google. | gemini-2.5-flash |
| Gemini 2.5 Pro | Advanced multimodal model by Google. | gemini-2.5-pro |
| Gemma 3 (27B) | 27-billion-parameter open model by Google. | gemma3:27b |
| GPT OSS (120B) | 120-billion-parameter open model by OpenAI. | gpt-oss:120b |
| GPT-4.1 | Enhanced GPT-4 model by OpenAI. | gpt-4.1-2025-04-14 |
| GPT-4o | Optimized GPT-4 model by OpenAI. | gpt-4o-2024-08-06 |
| GPT-5 | Advanced GPT-5 model by OpenAI. | gpt-5-2025-08-07 |
| GPT-5 Mini | Compact version of GPT-5 by OpenAI. | gpt-5-mini-2025-08-07 |
| Llama 3.3 (70B) | 70-billion-parameter model by Meta. | llama3.3:70b |
| Llama 4 (128B-A17B) | 128-billion-parameter mixture model by Meta. | llama4:128 × 17b |
| Mistral Small 3.1 (24B) | 24-billion-parameter model by Mistral AI. | mistral-small3.1:24b |
| Qwen 2.5 (72B) | 72-billion-parameter model by Alibaba. | qwen2.5:72b |
| Qwen 3 (235B) | 235-billion-parameter model by Alibaba. | qwen3:235b |
| Qwen 3 (235B-A22B) | 235-billion-parameter mixture model by Alibaba. | qwen3:235b-a22b |
| Qwen 3 (32B) | 32-billion-parameter model by Alibaba. | qwen3:32b |
| QwQ (32B) | 32-billion-parameter reasoning model by Alibaba. | qwq:32b |

GPT-4o demonstrated the most consistent and accurate performance in preserving narrative content while effectively removing identifying information.

The anonymization process employed a specialized prompt that instructed the LLM to remove or replace detective identifiers with neutral descriptors while preserving the original narrative structure and content. The complete anonymization prompt is presented in Prompt 1.

**Prompt 1.** *Original Story: {$S_{json}$}*
*Task: Anonymize the detective story provided above. The goal is to remove all direct references to the detective's name and gender, while keeping the story otherwise unchanged.*
*Follow these rules:*

- *Remove or replace all occurrences of the detective's name with indirect references, such as "the detective", "the investigator", or "the sleuth".*
- *Eliminate or neutralize all gendered terms and pronouns related to the detective (e.g., replace he/she with "they", his/her with "their" when needed for grammatical correctness).*
- *Apply the same logic for well-known characters or places directly related to the original detective (e.g., Watson and Baker Street in the case of Sherlock Holmes).*
- *Maintain the exact JSON structure of the input, including all keys, fields, and lists.*

*Return the fully anonymized story in the exact same JSON format as the original story, preserving the exact number of events as the original story. Do not include any explanations, summaries, or additional text. Return only the final JSON.*

Each of the 210 original stories was processed individually through this anonymization workflow. The resulting narratives, which we refer to as the *anonymized stories*, preserved all investigative content while removing only explicit detective identifiers, ensuring that subsequent identification would rely solely on the embedded investigative characteristics rather than naming conventions. Both the original and anonymized versions of all stories are included in the publicly available dataset, accessible through an interactive dataset explorer that enables side-by-side comparison of story versions: https://narrativelab.org/ai-sleuths-dataset/.

### 5.1.4. Detective identification

To assess whether the anonymized stories contained sufficient distinctive investigative characteristics to enable detective identification, we employed a multi-LLM evaluation approach. Given that the ability to recognize subtle investigative styles through narrative patterns is a complex interpretive task, we evaluated a set of 22 LLMs to identify which models demonstrated sufficient analytical capabilities for this task. The evaluated LLMs encompassed both open and proprietary models from multiple providers, including OpenAI, Anthropic, Google, Meta, Alibaba, DeepSeek, and Mistral, with varying parameter sizes and architectural designs. The complete list of models, including their descriptions and specific versions, is presented in Table 1.

The identification task employed a specialized prompt that instructed each LLM to analyze the anonymized story and identify which of the seven fictional detectives served as the protagonist based solely on their behavior, investigative method, and reasoning style as manifested in the narrative. The prompt explicitly constrained responses to the seven detectives included in this study and required both a detective selection and a justification for that selection. The complete detective identification prompt is presented in Prompt 2.

Proprietary models, such as those from OpenAI, Google, and Anthropic, were accessed via their respective APIs, while open models were executed locally using a self-hosted Ollama server.[4] Each model was queried independently using identical parameters to ensure uniform conditions across all responses. A temperature setting of 0.0 was applied to reduce variability and promote deterministic outputs, maximizing the consistency of model predictions.

**Prompt 2.** *Story: {$S_{title}$} {$S_{events}$}*
*Task: You are given a detective story in which the main character is a well-known fictional detective, but their name and gender have been removed. Your task is to identify which detective is the protagonist of the story, based entirely on their behavior, investigative method, and reasoning style as shown in the narrative.*
*Choose only from the following list:*

- *Hercule Poirot*
- *Sherlock Holmes*
- *William Murdoch*
- *Columbo*
- *Father Brown*
- *Miss Marple*
- *Auguste Dupin*

[4] https://ollama.com/.

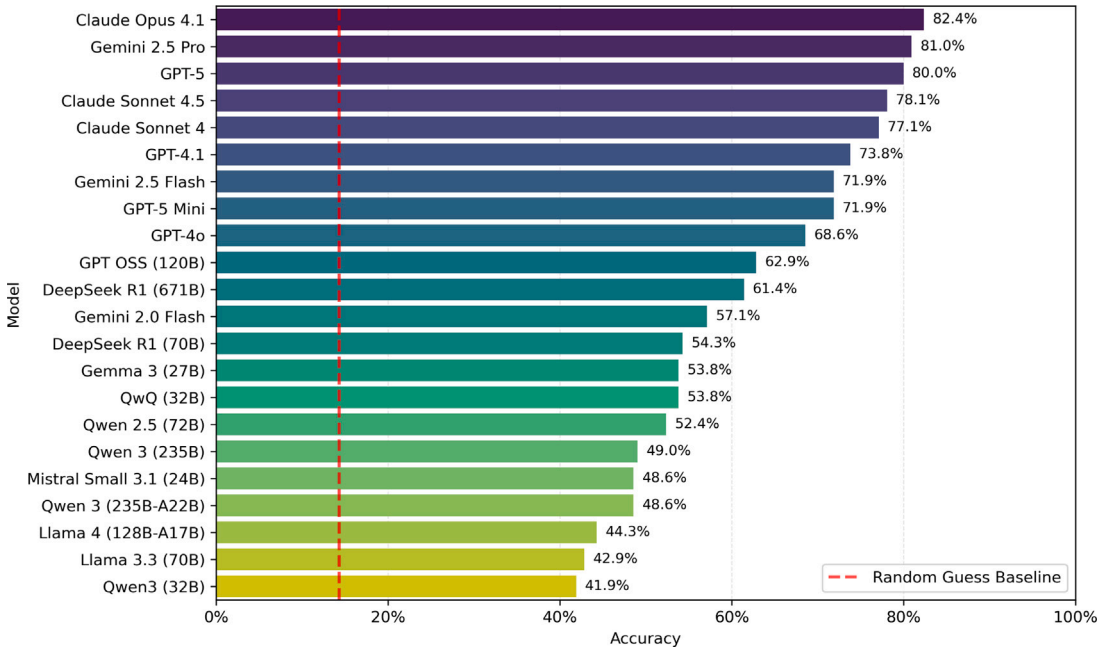


**Fig. 3.** Detective identification accuracy across the 22 evaluated LLMs. The vertical dashed line indicates the random guess baseline (14.3%).

*Output format:*

- *DETECTIVE: the name of the selected detective*
- *EXPLANATION: an explanation justifying the selection of the detective*

### 5.1.5. Results and discussion

The detective identification task revealed substantial variation in model performance across the 22 evaluated LLMs, with accuracy rates ranging from 41.9% to 82.4% (Fig. 3). Three state-of-the-art models achieved notably high accuracy: Claude Opus 4.1 (82.4%, 173 correct identifications out of 210 stories), Gemini 2.5 Pro (81.0%, 170/210), and GPT-5 (80.0%, 168/210). These top performers represent different model providers and architectural approaches, suggesting that the ability to recognize distinctive investigative styles generalizes across advanced LLMs rather than being specific to a particular architectural methodology. All models performed substantially above a random guess baseline of 14.3%, demonstrating that the anonymized stories retained identifiable characteristics of the detectives' investigative style. The fact that multiple state-of-the-art LLMs achieved approximately 80% accuracy indicates that the trait-guided generation method successfully produced narratives with sufficiently distinctive investigative characteristics of the selected detectives. The wide variation in model performance reflects differences in LLM analytical capabilities for this interpretive task rather than limitations in the distinctiveness of the generated stories.

Examining per-detective identification accuracy across all evaluated models reveals certain differences in the recognizability of different investigative styles (Fig. 4). Father Brown achieved consistently high identification rates across the top performing models, with all Claude models reaching perfect accuracy and several others exceeding 90%, a pattern consistent with his perfect identification in our preliminary trait validation study [13]. William Murdoch presented the most consistent identification challenge, with accuracy rarely exceeding 60% across nearly all models regardless of their overall performance capability. This difficulty mirrors the confusion observed in our preliminary study, where Murdoch's trait profile was occasionally misidentified as Sherlock Holmes, suggesting that the overlap between their scientific and analytical approaches persists when manifested in narrative form. Auguste Dupin exhibited notable variability in identification rates, ranging from 0% to 96.7% across different models. While our preliminary study reported only 53.3% accuracy for recognizability of Dupin's investigative methods due to frequent misclassification as Sherlock Holmes, the more advanced models used in the present study achieved substantially higher accuracy, suggesting that more sophisticated LLMs demonstrate improved capability to distinguish between these methodologically similar characters. The remaining detectives showed moderate to strong identification rates, consistent with their robust performance in trait profile identification reported in our preliminary study [13].

To assess whether the crime premise influenced detective identification accuracy, we conducted separate one-way Analyses of Variance (ANOVAs) for each model, with premise as the independent variable (3 levels) and identification correctness as the dependent variable (70 stories per premise for each model). The analysis revealed that premise had no significant effect on accuracy for the majority of models, with only 2 of 22 models showing significant differences: Claude Sonnet 4 ($F(2, 207) = 8.00$, $p < .001$) and Claude Sonnet 4.5 ($F(2, 207) = 3.96$, $p = .021$). Post-hoc Tukey HSD tests indicated that for both models, accuracy was significantly lower for Premise 2 (the Poivrière bar scenario) compared to Premise 1 (the Roman Theater mystery), while no significant differences emerged between Premises 1 and 3 or between Premises 2 and 3. Notably, Claude Opus 4.1, the best-performing model overall, showed no significant premise effect ($F(2, 207) = 0.42$, $p = .656$), with consistently high accuracy across all three premises (Premise 1: 85.7%, Premise 2: 80.0%, Premise 3: 81.4%). These findings demonstrate that for most models, including the top performers, the distinctiveness of detective investigative styles remained recognizable regardless of the crime premise, suggesting that the trait-guided generation method successfully produces detective-specific narratives independently of crime premise and plot context.

Given Claude Opus 4.1's superior performance, we conducted a detailed analysis of its identification patterns to examine how detective styles manifested in the generated narratives. The model achieved 82.4% overall accuracy with balanced macro-averaged performance: precision of 0.856, recall of 0.824, and F1-score of 0.824. The complete per-detective classification metrics are presented in Table 2, and the confusion matrix is shown in Table 3.

The classification metrics reveal substantial variation in detective recognizability. Father Brown and Columbo demonstrated the strongest performance: Father Brown achieved perfect recall (1.00) with all

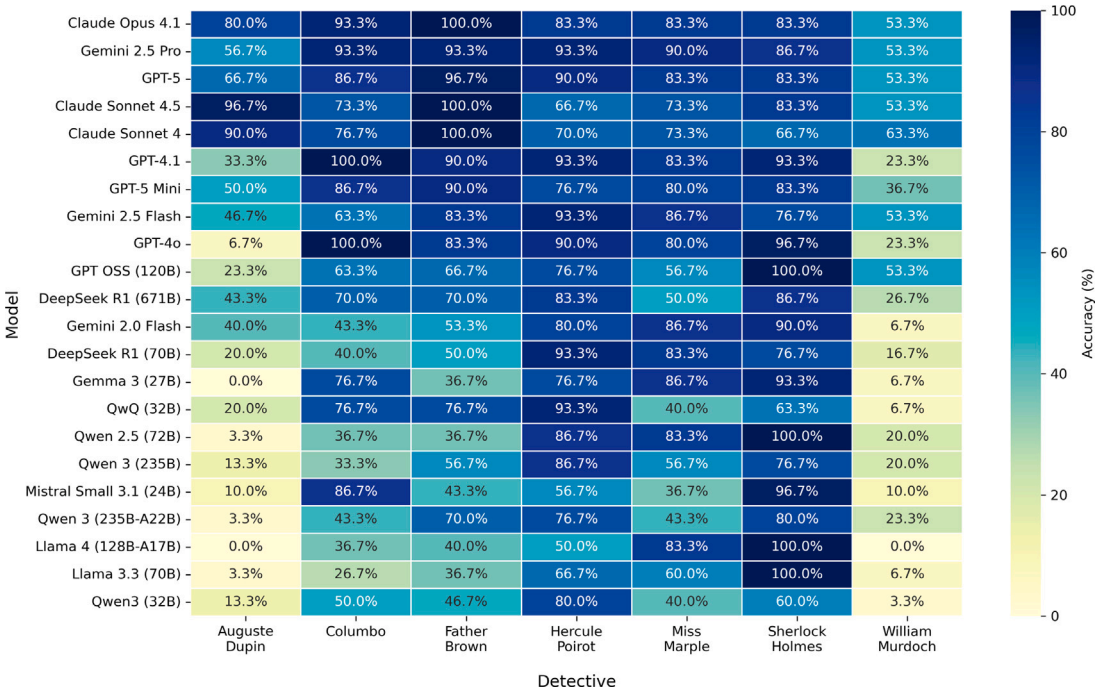


**Fig. 4.** Per-detective identification accuracy heatmap across all 22 evaluated LLMs. Rows represent models (sorted by overall performance) and columns represent detectives. Color intensity indicates accuracy percentage, with darker blue showing higher accuracy.

**Table 2**
Classification metrics per detective for Claude Opus 4.1: precision indicates reliability when predicting a detective, recall indicates the proportion of stories correctly identified, and F1-score is the harmonic mean of precision and recall.

| Detective | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Auguste Dupin | 0.73 | 0.80 | 0.76 | 30 |
| Columbo | 1.00 | 0.93 | 0.97 | 30 |
| Father Brown | 0.75 | 1.00 | 0.86 | 30 |
| Hercule Poirot | 0.89 | 0.83 | 0.86 | 30 |
| Miss Marple | 1.00 | 0.83 | 0.91 | 30 |
| Sherlock Holmes | 0.62 | 0.83 | 0.71 | 30 |
| William Murdoch | 1.00 | 0.53 | 0.70 | 30 |
| **Macro avg** | **0.86** | **0.82** | **0.82** | **210** |

stories correctly identified, while Columbo achieved the highest overall F1-score (0.97) by combining near-perfect recall (0.93) with perfect precision (1.00). Three detectives achieved perfect precision – Columbo, Miss Marple, and William Murdoch – indicating that predictions for these detectives were always correct when made, though their recall varied substantially (Columbo: 0.93, Miss Marple: 0.83, Murdoch: 0.53).

The confusion matrix also reveals interesting misidentification patterns. William Murdoch presented the most pronounced confusion pattern, with 14 of 30 stories misidentified: 9 attributed to Sherlock Holmes, 3 to Auguste Dupin, and single instances to Hercule Poirot and Father Brown. This asymmetric pattern, with Murdoch frequently mistaken for others but never vice versa, indicates that his scientific and forensic methodology overlaps substantially with other analytically-oriented detectives. The perfect precision (1.00) demonstrates that when Murdoch's distinctive characteristics are recognized, they are highly reliable identifiers, but these markers may not manifest consistently across all generated narratives.

Auguste Dupin and Sherlock Holmes exhibited a bidirectional confusion, with 5 Dupin stories misidentified as Holmes and 3 Holmes stories as Dupin, reflecting the well-documented methodological similarities between these characters [40]. This mutual confusion also impacted Holmes's precision (0.62), as stories from multiple other detectives were occasionally attributed to him, suggesting his analytical style may be perceived as a general investigative archetype. Additionally, several stories featuring Hercule Poirot, Miss Marple, and Sherlock Holmes were misattributed to Father Brown (3, 4, and 1, respectively), contributing to Father Brown's lower precision (0.75) despite perfect recall.

To determine whether the identification patterns observed in Claude Opus 4.1 generalized across other high-performing LLMs, we assessed inter-rater agreement among the top three models (treating each model as an independent rater) using Fleiss' Kappa. The analysis revealed substantial agreement ($\kappa = 0.78$), indicating that these models converged on similar identification patterns despite their different architectures and training methodologies. Pairwise Cohen's Kappa values ranged from 0.75 to 0.80, with Claude Opus 4.1 and GPT-5 showing the strongest agreement ($\kappa = 0.80$, 83.3% simple agreement). The three models reached unanimous consensus on 154 of 210 stories (73.3%), disagreed partially on 48 stories (22.9%), and completely diverged on only 8 stories (3.8%). Notably, when all three models agreed on an identification, their consensus accuracy was 94.8% (146/154 correct), substantially higher than their individual accuracies. This pattern demonstrates that model consensus serves as a reliable indicator of identification confidence, and the high inter-rater agreement across diverse architectures provides strong evidence that the trait-guided generation method produces narratives with objectively recognizable detective-specific investigative characteristics rather than patterns that are artifacts of particular model biases.

It should be noted that while the anonymization process effectively removes explicit detective identifiers, certain narrative elements may persist, such as the presence of a known companion character, distinctive physical characteristics of the detective, or characteristic linguistic styles, which could remain diagnostic of particular detectives independently of their investigative methods. To further investigate the factors contributing to successful identifications, we performed a qualitative analysis of all 210 explanations provided by Claude Opus 4.1 for its detective identifications. Each explanation was systematically reviewed and manually categorized according to its primary reasoning

**Table 3**
Confusion matrix for Claude Opus 4.1, where rows represent actual detectives and columns represent predicted detectives.

| Actual/Predicted | A. Dupin | Columbo | F. Brown | H. Poirot | M. Marple | S. Holmes | W. Murdoch |
|---|---|---|---|---|---|---|---|
| Auguste Dupin | 24 | 0 | 1 | 0 | 0 | 5 | 0 |
| Columbo | 0 | 28 | 0 | 1 | 0 | 1 | 0 |
| Father Brown | 0 | 0 | 30 | 0 | 0 | 0 | 0 |
| Hercule Poirot | 2 | 0 | 3 | 25 | 0 | 0 | 0 |
| Miss Marple | 1 | 0 | 4 | 0 | 25 | 0 | 0 |
| Sherlock Holmes | 3 | 0 | 1 | 1 | 0 | 25 | 0 |
| William Murdoch | 3 | 0 | 1 | 1 | 0 | 9 | 16 |

basis: (1) Investigative Method & Approach; (2) Linguistic Features & Speech Patterns; (3) Physical Appearance & Behavioral Characteristics; or (4) Associated Characters & Setting/Location/Culture Cues. When multiple reasoning elements were present, categorization was based on the element that received the most explicit emphasis in the explanation. The analysis revealed that 56.2% (118/210) of identifications were based primarily on investigative methods, while 43.8% (92/210) relied on additional character cues embedded in the narrative context. A chi-square test of independence confirmed a significant association between reasoning theme and identification correctness ($\chi^2 = 13.77$, $p = .003$). Investigative methods were the predominant reasoning basis for both correct (53.8%) and incorrect (67.6%) identifications, with the latter reflecting cases where overlapping investigative characteristics between detectives, such as the analytical approaches shared by Auguste Dupin and Sherlock Holmes, led to confusion. While contextual cues appeared more frequently in incorrect identifications (29.7%) than correct ones (15.6%), most identifications involving such elements were still accurate (80 of 92 cases, 87.0%), indicating that contextual narrative elements generally facilitated identification despite occasionally contributing to misclassification.

The prevalence of contextual narrative cues varied substantially by detective. Hercule Poirot (66.7%), Sherlock Holmes (63.3%), and Father Brown (60.0%) showed the highest reliance on such cues, while Columbo and Auguste Dupin exhibited the lowest (26.7% each). Specific patterns emerged for individual detectives: Hercule Poirot identifications frequently involved linguistic features, particularly French phrases (50.0% of cases), while Sherlock Holmes identifications often referenced associated characters or locations (50.0%). Father Brown's physical appearance and behavioral characteristics, particularly clerical attire and demeanor, featured prominently in explanations (56.7%). These patterns indicate that certain detectives possess more distinctive contextual markers that naturally appear in narratives where they are protagonists.

To illustrate how investigative characteristics and contextual elements manifest in the explanations, consider two representative cases from Claude Opus 4.1: one where identification relied solely on investigative methods and another where contextual narrative elements were also identified by the model. In the first case, Sherlock Holmes was identified based purely on investigative characteristics. The model's explanation cited "methodical physical examination of evidence", noting that "the detective meticulously examines physical clues — the padlock, the snow, the pocket watch's internal mechanism", alongside "disguise usage" for undercover investigation and "mechanical and technical knowledge", observing that "the detective understands the pocket watch's internal mechanism, discovering a rigged switch inside". These patterns align with Holmes's synthesized traits of observing minute details, utilizing scientific knowledge, and employing undercover strategies. In the second case, Hercule Poirot was identified through a combination of investigative and contextual elements. The explanation noted investigative characteristics such as "methodical scene examination" where "the detective carefully examines every detail of the crime scene systematically" and "psychological insight" focusing on "understanding motivations and human nature", but also cited contextual cues including "formal, polite speech patterns" with "formal French phrases like 'Messieurs,' 'Merci, mademoiselle,' and 'monsieur'". While the linguistic markers represent additional narrative clues beyond investigative method, the methodical and psychological characteristics correspond to Poirot's synthesized investigative traits. Although the contextual elements facilitated identification, genuine investigative characteristics remained present in the generated narratives.

When restricting analysis to identifications based solely on investigative methods (118 cases), the adjusted overall accuracy was 78.8%, compared to the original 82.4%. Per-detective adjusted accuracy revealed additional variations (Fig. 5): Father Brown maintained perfect accuracy (100%), while Columbo and Auguste Dupin achieved over 90%. In contrast, William Murdoch and Hercule Poirot showed notably lower accuracy when stories containing additional character clues were excluded (45% and 60%, respectively), suggesting that their investigative style, when isolated from contextual cues, proved less distinctive in the generated stories.

To further illustrate how investigative traits manifest in the generated narratives, we present two representative story excerpts that demonstrate the direct translation of trait profiles into narrative content, contrasting two markedly different investigative styles. In a Columbo story generated for Premise 3, the traits of feigning forgetfulness and uncovering contradictions through persistent questioning can be observed in the following exchange:

> *"Finn", Columbo said, "you ever go into the church after midnight?" Finn blinked. "No. I mean... I used to dream about it. But I never went". Columbo smiled. "I see. You know, the bell rang once last night. You were the only one who could've done it". Finn looked away. "I don't even know how the bell works". Columbo leaned forward. "Ah. But you used to know, didn't you?" Finn's face went pale.*

In contrast, a Father Brown story generated for Premise 2 reflects an entirely different investigative approach, centered on moral motivation and psychological understanding rather than physical evidence or strategic questioning:

> *"It is not the how of the crime that interests me", he muses. "It is the why". [...] The pieces fall into place. The soldier, a man with a past of violence and regret, had been drawn to Lacheneur under the pretense of revenge. But Lacheneur was not the enemy. He was the mirror. The soldier had come to provoke, to test, to find someone to blame for his own fall. [...] "You did not kill him", Father Brown says. "You tried to protect yourself. But you did not kill him". [...] "You did what you thought you had to. But now you know the truth".*

These excerpts illustrate how the trait-guided method produces narratives that are not only stylistically distinct but also internally consistent with each detective's synthesized trait profile: Columbo's disarming demeanor and strategic use of feigned uncertainty contrast sharply with Father Brown's empathetic reconstruction of the criminal's moral fall and his focus on the why of a crime rather than the how, reflecting the specific traits of each detective.

These findings demonstrate that the trait-guided generation method successfully produced detective-specific narratives with recognizable investigative characteristics. When restricting to identifications based solely on investigative methods, the overall accuracy remained high at

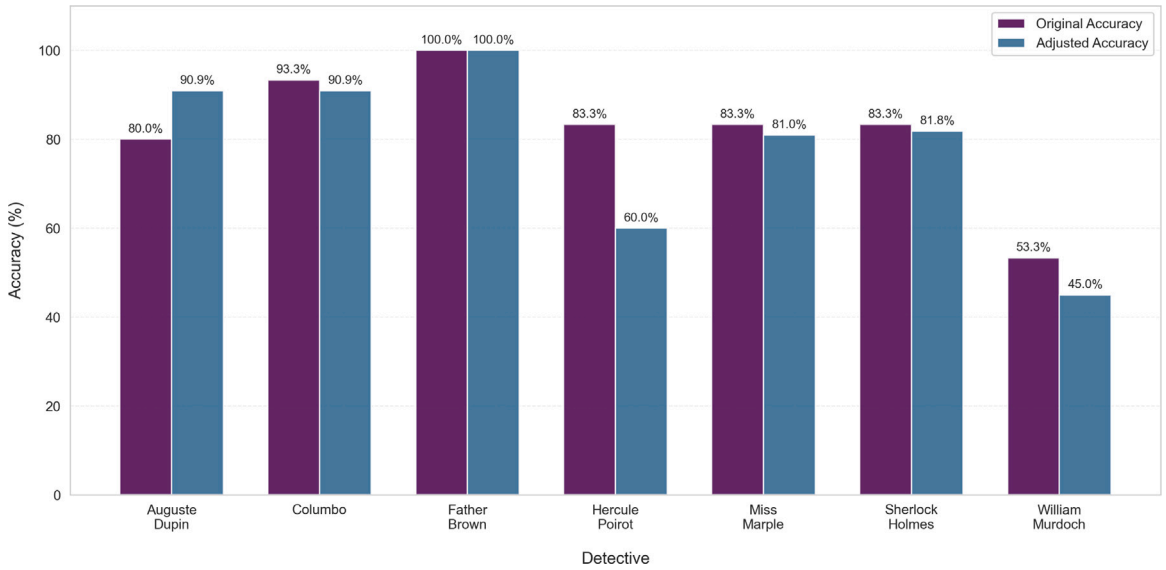


**Fig. 5.** Comparison of original and adjusted detective identification accuracy by detective for Claude Opus 4.1. Adjusted accuracy excludes identifications based on contextual narrative elements.

78.8% (compared to 82.4% with all cues), and several detectives maintained strong identification rates. While contextual narrative elements contributed to the identification for some detectives, particularly those with distinctive linguistic or cultural markers such as Hercule Poirot and Sherlock Holmes, the minimal decline in accuracy when excluding these elements suggests the effectiveness of the trait-guided generation method in embedding distinctive investigative styles in the generated narratives.

### 5.2. Phase 2: Story quality

In the second phase of this study, we assessed the narrative quality of the generated stories using a LLM-based evaluation method. This approach builds on recent research validating LLMs as effective evaluators for narrative quality that aligns with human judgment. Chhun et al. [41] established a framework of six orthogonal evaluation criteria for story quality (relevance, coherence, empathy, surprise, engagement, and complexity), grounded in social sciences literature on storytelling effectiveness. Subsequently, Chhun et al. [42] demonstrated that LLMs, when prompted with these criteria, achieve strong correlations with human judgment, outperforming traditional automatic metrics. Given the inherent subjectivity of story quality evaluation, where individual preferences for narrative elements vary significantly across readers, and the demonstrated reliability of LLM-based evaluation as a proxy for human assessment, we employed this validated methodology to compare the narrative quality of trait-guided stories against baseline stories generated without investigative trait guidance.

#### 5.2.1. Baseline generation

To isolate the effect of trait guidance on narrative quality, we generated a prompt ablation baseline dataset of 210 stories using the same detective-premise combinations employed in Phase 1 (7 detectives × 3 premises × 10 stories). The ablation baseline omits the investigative trait profiles and narrative guidance principles that comprise the trait-guided method presented in Section 4, providing only the detective name and crime premise as inputs. The prompt instructed the LLM to write a detective story featuring the specified detective investigating and solving the given crime, requiring the same structured output format as the trait-guided stories (title followed by 12 to 18 numbered narrative events with scene and image descriptions) to ensure comparability between conditions. The complete baseline story generation prompt is presented in Appendix F.

All baseline stories were generated using the same LLM employed for the trait-guided generation (Qwen 3 32B) to isolate the effect of the trait-guided method from potential model-specific variations. Trait-guided stories were on average longer than baseline stories ($M$ = 1066 words, $SD$ = 323, vs. $M$ = 748 words, $SD$ = 125), with trait-guided events also averaging more words per event ($M$ = 78.72, $SD$ = 30.70, vs. $M$ = 48.77, $SD$ = 12.18 for baseline events). This difference is expected given the nature of the trait-guided approach, which directly instructs the LLM to use the investigative trait profiles to produce more elaborated narratives compared to the simplified baseline prompt. The potential influence of story length on quality scores is assessed through length-controlled analyses reported in Section 5.2.3. Both the trait-guided and baseline story datasets are included in the publicly available dataset at: https://narrativelab.org/ai-sleuths-dataset/.

#### 5.2.2. Multi-LLM story evaluation

To evaluate the narrative quality of both trait-guided and baseline stories, we employed the three best-performing models from Phase 1 of our study: Claude Opus 4.1, Gemini 2.5 Pro, and GPT-5. These models were selected based on their demonstrated analytical capabilities in the detective identification task, where they achieved approximately 80% accuracy in recognizing investigative styles from narrative patterns. Additionally, these models represent diverse architectural approaches from different providers (Anthropic, Google, and OpenAI, respectively), ensuring that the evaluation of story quality is not dependent on a single model's characteristics or training methodology.

Following the evaluation framework established by Chhun et al. [41,42], each story was evaluated across all their six above mentioned criteria: relevance, coherence, empathy, surprise, engagement, and complexity. For each criterion, models were instructed to critically evaluate the story using detailed rating guidelines adapted from Chhun et al.'s annotation protocol and provide both a rating on a 5-point Likert scale and a justification citing specific story elements. Each criterion was evaluated independently through a separate prompt that included the crime premise, the complete story text, the criterion definition, detailed rating scale guidelines, and instructions for the output format. The evaluation prompt structure and complete rating guidelines for all six criteria are presented in Appendix G.

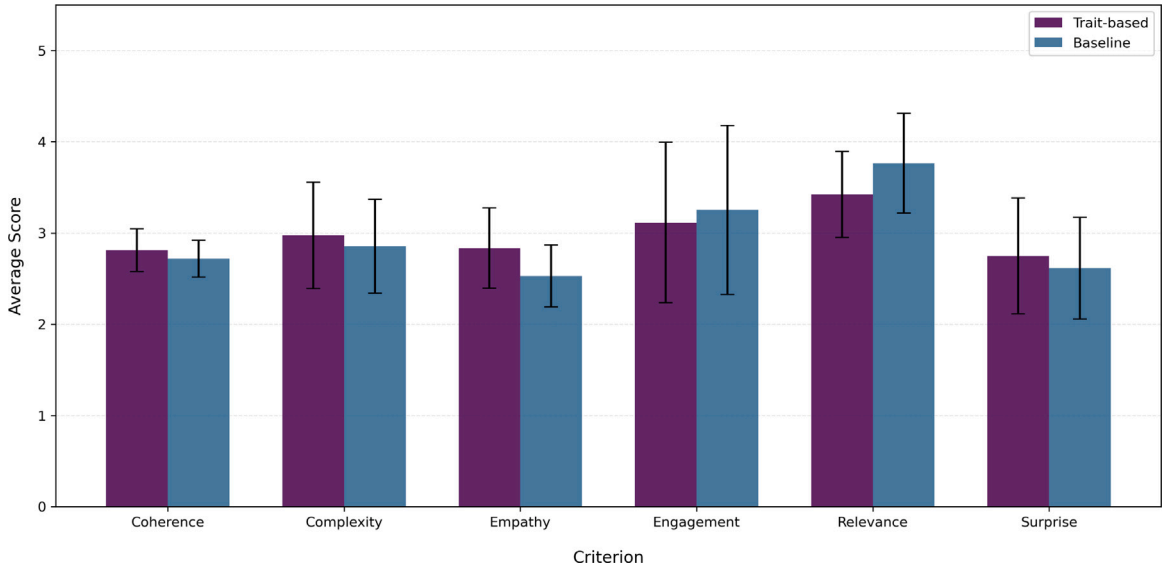


**Fig. 6.** Average story quality scores by criterion for trait-guided and baseline stories (averaged across Claude Opus 4.1, Gemini 2.5 Pro, and GPT-5). Error bars represent standard deviation across evaluator models.

The three LLMs were accessed via their respective APIs (Anthropic API[5] for Claude Opus 4.1, Google AI API[6] for Gemini 2.5 Pro, and OpenAI API[7] for GPT-5). To maximize consistency and reduce variability in the evaluation process, a temperature setting of 0.0 was applied across all models. Each of the 420 stories (210 trait-guided and 210 baseline) was independently evaluated by all three models across all six criteria, resulting in 7560 individual assessments (420 stories × 3 models × 6 criteria). This evaluation design enabled comparative analysis of narrative quality between the two generation approaches, as well as analysis of inter-rater agreement among the evaluator models.

### 5.2.3. Results and discussion

The LLM-based evaluation revealed that trait-guided stories achieved higher average scores on four of the six quality criteria when compared to the simplified prompt baseline (Fig. 6). Trait-guided stories showed higher scores in coherence (2.81 vs. 2.72), complexity (2.97 vs. 2.86), empathy (2.83 vs. 2.53), and surprise (2.75 vs. 2.61), with all three LLM evaluators showing consistent directional agreement on these dimensions. Baseline stories achieved higher scores on relevance (3.76 vs. 3.42), while engagement showed mixed results across models (3.11 vs. 3.25). The rating guidelines provided in Appendix G, enable direct interpretation of these score values in the context of each criterion.

Paired t-tests conducted for each LLM individually confirmed the consistency of the patterns observed in the overall average scores while revealing some variation in statistical significance across criteria (Table 4). Empathy demonstrated the most robust effect, with all three models showing statistically significant improvements for trait-guided stories (GPT-5: $t = 5.81$, $p < .001$, $d = 0.40$; Claude Opus 4.1: $t = 5.42$, $p < .001$, $d = 0.37$; Gemini 2.5 Pro: $t = 6.08$, $p < .001$, $d = 0.42$). Similarly, all three models showed significantly lower relevance scores for trait-guided stories (GPT-5: $t = -6.93$, $p < .001$, $d = -0.48$; Claude Opus 4.1: $t = -5.11$, $p < .001$, $d = -0.35$; Gemini 2.5 Pro: $t = -4.30$, $p < .001$, $d = -0.30$). For coherence, surprise, and complexity, GPT-5 and Gemini 2.5 Pro showed significant advantages for trait-guided stories on at least one criterion each, while Claude Opus 4.1 showed no significant differences on these dimensions. Engagement is the only criterion where trait-guided stories scored significantly lower, though this effect appeared only in Claude Opus 4.1 ($t = -4.47$, $p < .001$, $d = -0.31$) and Gemini 2.5 Pro ($t = -2.92$, $p = .004$, $d = -0.20$), while GPT-5 showed no significant difference.

**Table 4**
Paired t-test results comparing trait-guided and baseline stories for each LLM evaluator. Positive differences indicate higher scores for trait-guided stories. Significant results ($p < .05$) are shown in bold.

| Model | Criterion | Mean Diff | $t$ | $p$ | Cohen's $d$ |
|---|---|---|---|---|---|
| GPT-5 | Relevance | **−0.41** | **−6.93** | **<.001** | **−0.48** |
| | Coherence | **0.13** | **3.48** | **<.001** | **0.24** |
| | Empathy | **0.24** | **5.81** | **<.001** | **0.40** |
| | Surprise | **0.21** | **3.91** | **<.001** | **0.27** |
| | Engagement | 0.02 | 0.82 | .416 | 0.06 |
| | Complexity | **0.10** | **2.45** | **.015** | **0.17** |
| Claude Opus 4.1 | Relevance | **−0.25** | **−5.11** | **<.001** | **−0.35** |
| | Coherence | 0.08 | 1.64 | .103 | 0.11 |
| | Empathy | **0.20** | **5.42** | **<.001** | **0.37** |
| | Surprise | 0.03 | 1.51 | .134 | 0.10 |
| | Engagement | **−0.19** | **−4.47** | **<.001** | **−0.31** |
| | Complexity | 0.06 | 1.40 | .164 | 0.10 |
| Gemini 2.5 Pro | Relevance | **−0.36** | **−4.30** | **<.001** | **−0.30** |
| | Coherence | 0.08 | 1.12 | .266 | 0.08 |
| | Empathy | **0.48** | **6.08** | **<.001** | **0.42** |
| | Surprise | 0.17 | 1.84 | .068 | 0.13 |
| | Engagement | **−0.25** | **−2.92** | **.004** | **−0.20** |
| | Complexity | **0.20** | **2.41** | **.017** | **0.17** |

Given the observed length difference between conditions, we conducted additional analyses to assess whether story length constituted a confounding variable in the quality score comparisons. An ANCOVA with word count as a covariate confirmed that the condition effects on relevance, coherence, empathy, and engagement remained significant after covariate adjustment (relevance: $F = 54.95$, $p < .001$; coherence: $F = 53.35$, $p < .001$; empathy: $F = 9.56$, $p = .002$; engagement: $F = 20.11$, $p < .001$). For surprise and complexity, however, the condition effect was no longer significant after controlling for word

[5] https://docs.claude.com/en/api/overview.
[6] https://cloud.google.com/vertex-ai/generative-ai/docs.
[7] https://platform.openai.com/docs/api-reference.

count (surprise: $F = 0.03$, $p = .863$; complexity: $F = 0.12$, $p = .731$), indicating that the score differences observed for these two criteria may reflect the greater narrative detail and length produced by the trait-guided approach rather than the specific influence of investigative trait profiles on these narrative dimensions. This interpretation is consistent with the definitions of these criteria: complexity explicitly measures how elaborate and well crafted story elements are, making it inherently sensitive to narrative length, while surprise depends on the distribution of clues across the narrative, which can be directly affected by story length. Partial correlation analysis revealed a particularly noteworthy pattern for coherence: once length was controlled, the condition effect increased substantially (zero-order correlation $r = 0.08$ vs. partial correlation $r = 0.20$), indicating that the raw scores underestimated the true advantage of trait-guided stories on this criterion. This suppression occurs because longer stories are expected to be inherently more challenging to keep internally consistent, which attenuated the coherence advantage of trait-guided stories in the unadjusted analysis. That trait-guided stories maintain stronger narrative coherence despite being substantially longer suggests that the structured investigative guidance contributes meaningfully to logical consistency and narrative flow.

Per-evaluator analysis revealed meaningful differences in sensitivity to story length across the three evaluator models. Gemini 2.5 Pro showed the strongest length sensitivity, with the association between condition and score decreasing substantially after controlling for word count for engagement (zero-order $r = -0.14$ vs. partial $r = -0.34$) and complexity (zero-order $r = 0.13$ vs. partial $r = -0.14$), and a non-significant condition effect on empathy once length was controlled. GPT-5 demonstrated the most stable condition effects, with most criteria remaining significant after controlling for word count, including surprise ($F = 7.15$, $p = .008$), which was no longer significant in the overall analysis across all three evaluator models. Claude Opus 4.1 showed an intermediate pattern, with condition effects generally remaining significant after controlling for word count across most criteria. These differences in length sensitivity across evaluator models suggest that part of the variability in model ratings may reflect differential implicit weighting of story length and detail when assessing narrative quality.

To further assess the reliability of the evaluation and examine the sources of inter-evaluator variability, we calculated inter-rater agreement among the three LLM evaluators using the Intraclass Correlation Coefficient (ICC). The ICC values ranged from 0.03 to 0.31 across the six criteria (Table 5), indicating low to moderate agreement among the models. Empathy achieved the highest agreement (ICC = 0.31), followed by coherence (ICC = 0.25) and relevance (ICC = 0.16), while engagement (ICC = 0.03), surprise (ICC = 0.08), and complexity (ICC = 0.08) showed minimal agreement. These values align with patterns observed in human evaluation of story quality. Chhun et al. [41] reported ICC values ranging from 0.29 to 0.56 for human evaluators, with higher agreement on complexity (0.56) and lower agreement on criteria such as surprise (0.28) and coherence (0.29). The similar pattern of variable agreement across criteria, with certain dimensions permitting more consistent evaluation than others, demonstrates that the subjectivity inherent in narrative quality assessment affects both humans and LLMs.

To assess the correspondence between the LLM-based evaluation and human judgment in the context of our detective stories, we conducted a preliminary validation study using a balanced subset of 42 stories drawn from the full dataset (7 detectives × 3 premises × 1 trait-guided story and 1 baseline story per detective-premise combination). Three human evaluators, recruited via convenience sampling and without prior expertise in detective fiction or literary analysis, independently evaluated each story across the same six criteria using identical rating guidelines and Likert scale as the LLM evaluation. Stories were presented without revealing the generation condition, and the order was randomized across evaluators. Comparing human and LLM scores averaged across evaluators, empathy emerged as the most strongly validated criterion, with substantial human-LLM agreement (Pearson $r = 0.63$, $p < .001$; ICC(A,1) = 0.58, 95% CI [0.34, 0.75]). Each of the three LLM evaluators individually correlated strongly with human judgments on empathy (Pearson $r = 0.50 - 0.59$, all $p < .001$), indicating that this agreement is not driven by any single model. Coherence showed modest but significant agreement (Pearson $r = 0.31$, $p = .047$), while the remaining criteria exhibited weaker absolute agreement (see Table 6), consistent with the subjectivity inherent in narrative quality evaluation noted by Chhun et al. [41] and reflecting the low inter-rater agreement observed for these dimensions in both human and LLM groups.

**Table 5**
Inter-rater agreement (ICC) among the three evaluator models for each quality criterion.

| Criterion | ICC | 95% CI |
|---|---|---|
| Relevance | 0.16 | [0.05, 0.26] |
| Coherence | 0.25 | [0.17, 0.34] |
| Empathy | 0.31 | [0.13, 0.46] |
| Surprise | 0.08 | [0.00, 0.16] |
| Engagement | 0.03 | [−0.01, 0.08] |
| Complexity | 0.08 | [0.00, 0.16] |

**Table 6**
Human-LLM agreement on the 42-story validation subset. Pearson $r$ measures rank consistency, while ICC(A,1) measures absolute agreement, penalizing systematic offset between evaluator types. Significant results ($p < .05$) shown in bold.

| Criterion | Pearson $r$ | $p$ | ICC(A,1) | 95% CI |
|---|---|---|---|---|
| Relevance | **0.38** | **.013** | **0.26** | [−0.05, 0.53] |
| Coherence | **0.31** | **.047** | **0.19** | [−0.07, 0.45] |
| Empathy | **0.63** | **<.001** | **0.58** | **[0.34, 0.75]** |
| Surprise | 0.26 | .092 | 0.19 | [−0.07, 0.44] |
| Engagement | 0.16 | .307 | 0.13 | [−0.13, 0.39] |
| Complexity | 0.30 | .057 | **0.27** | [−0.01, 0.52] |

Beyond pairwise correlations, we examined whether human evaluators reached the same verdict as the LLM evaluators on the trait-guided versus baseline comparison, using paired t-tests on the 21 detective-premise pairs (Table 7). The strongest convergence appeared on empathy, where both evaluator types found a significant trait-guided advantage, with 81% of detective-premise pairs favoring trait-guided stories in the human ratings. This convergence directly corroborates the most prominent finding of the LLM-based evaluation, although the per-pair pattern indicates that the LLM advantage is more concentrated than the human one (43% vs. 81% of pairs favoring trait-guided). On surprise and engagement, human evaluators showed a stronger trait-guided preference than the LLM evaluators, while the LLM evaluators showed no significant difference on surprise and a slight baseline preference on engagement. The human-rated trait-guided preference on these criteria is broad-based, with 81% of pairs favoring trait-guided stories on surprise and 62% on engagement, compared to 24% in the LLM ratings for both criteria. These patterns indicate that, where human and LLM verdicts diverge, the LLM evaluation appears more conservative than human judgment rather than overstating the trait-guided method's benefits.

These validation findings reinforce the empathy improvement observed in the LLM-based evaluation. The consistent improvement across all three LLM evaluators, with medium to large effect sizes ($d = 0.37$–$0.42$), demonstrates that the trait-guided method successfully enhances character depth and emotional portrayal within the generated stories. This outcome aligns with the nature of the investigative trait profiles, which emphasize psychological insight, understanding of human nature, and behavioral observation, as these characteristics naturally

**Table 7**
Paired t-test results comparing trait-guided and baseline stories on the 21 detective-premise pairs of the validation subset, computed independently for human and LLM evaluators. The % pairs columns indicate the proportion of pairs where trait-guided stories received higher mean scores than baseline stories. Significant results ($p < .05$) shown in bold.

| Criterion | Human evaluators | | | | LLM evaluators | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean Diff | $t$ | $p$ | % pairs | Mean Diff | $t$ | $p$ | % pairs |
| Relevance | +0.05 | +0.42 | .679 | 33.3% | −0.27 | −1.78 | .091 | 28.6% |
| Coherence | +0.03 | +0.30 | .766 | 33.3% | +0.19 | +1.50 | .150 | 42.9% |
| Empathy | **+0.79** | **+6.05** | **<.001** | 81.0% | **+0.24** | **+2.25** | **.036** | 42.9% |
| Surprise | **+0.49** | **+3.69** | **.001** | 81.0% | −0.11 | −1.10 | .285 | 23.8% |
| Engagement | **+0.44** | **+3.21** | **.004** | 61.9% | **−0.29** | **−2.34** | **.030** | 23.8% |
| Complexity | **+0.32** | **+2.35** | **.029** | 42.9% | −0.02 | −0.14 | .888 | 38.1% |

**Table 8**
Empathy scores by detective and generation method (averaged across all three evaluator models). Positive Cohen's $d$ values indicate higher empathy scores for trait-guided stories. Significant results ($p < .05$) shown in bold.

| Detective | Trait-Guided | Baseline | Cohen's $d$ | p-value |
|---|---|---|---|---|
| Father Brown | 3.56 | 2.68 | **1.25** | **<.001** |
| Miss Marple | 2.82 | 2.49 | **0.55** | **<.001** |
| Hercule Poirot | 2.82 | 2.52 | **0.44** | **.003** |
| Sherlock Holmes | 2.71 | 2.43 | **0.43** | **.004** |
| Columbo | 2.87 | 2.57 | **0.41** | **.007** |
| William Murdoch | 2.67 | 2.53 | 0.22 | .139 |
| Auguste Dupin | 2.40 | 2.48 | −0.12 | .407 |

translate into richer character development. The improvements in coherence, surprise, and complexity, while not universally significant across all models, showed consistent directional agreement among the LLM evaluators, with the human validation indicating that the trait-guided method's benefits on surprise may be more pronounced than the LLM scores alone suggest.

Within the full LLM evaluation, we further examined whether the effectiveness of trait guidance varied across different investigative styles by conducting two-way ANOVAs testing the interaction between generation approach (trait-guided vs. baseline) and detective for each quality criterion. A significant interaction emerged for empathy ($F(6, 1246) = 8.74$, $p < .001$), indicating that the impact of trait guidance on emotional characterization differed substantially across detectives. Per-detective analysis of empathy scores (Table 8) revealed particularly pronounced improvements for detectives characterized by psychological insight in their trait profiles. Father Brown stories demonstrated the largest enhancement under trait guidance (trait-guided: 3.56 vs. baseline: 2.68, Cohen's $d = 1.25$, $p < .001$), a substantial effect reflecting his trait profile's emphasis on intuitive understanding of human nature and moral reasoning. Similarly significant improvements appeared for Miss Marple ($d = 0.55$, $p < .001$), Hercule Poirot ($d = 0.44$, $p = .003$), Columbo ($d = 0.41$, $p = .007$), and Sherlock Holmes ($d = 0.43$, $p = .004$). In contrast, William Murdoch and Auguste Dupin showed no significant empathy improvements, consistent with their trait profiles' stronger emphasis on scientific methodology and analytical reasoning rather than psychological characterization. These patterns demonstrate that the trait-guided method's impact on narrative quality is not uniform but rather reflects the specific characteristics embedded in each detective's investigative profile, with psychologically-oriented detectives benefiting most in terms of emotional depth and character development.

To illustrate how these empathy improvements manifest in the generated narratives, we present representative evaluation explanations from Claude Opus 4.1 for a Father Brown story generated for Premise 2. The trait-guided story received a rating of 4, with the following justification: *"The story creates strong emotional resonance through Jean Lacheneur's desperate need to matter and be remembered rather than dismissed as a coward, culminating in his breakdown ('I didn't want to kill him. I just... I wanted to matter'). Father Brown's empathetic understanding of Lacheneur's 'soul in torment' and his recognition that the case was about 'who had the need to be seen' demonstrates sophisticated emotional insight"*. In contrast, the corresponding baseline story received a rating of 2: *"The story presents characters with basic emotions – the dying soldier's desperation, Lacheneur's arrogance, the young killer's panic – but these remain surface-level and underdeveloped. While Father Brown reflects on 'the weight of war, guilt, and the shadows of the past,' this philosophical musing is told rather than shown through genuine emotional depth. The characters function primarily as plot devices in a mystery framework rather than as emotionally complex individuals whose inner lives readers can meaningfully connect with"*. These contrasting explanations demonstrate that the evaluator's ratings were grounded in specific narrative elements reflecting the detective's investigative traits, as the trait-guided story's higher empathy score is explicitly attributed to Father Brown's characteristic empathetic reconstruction of the criminal's psychological state, a core element of his synthesized trait profile, rather than being an arbitrary numerical assessment.

While empathy demonstrated clear benefits from the trait-guided generation approach, the relevance criterion showed a consistent pattern of lower scores for trait-guided stories across all detectives. This uniform effect reflects an expected trade-off inherent in the method. The incorporation of investigative characteristics and narrative guidance principles introduces additional elements beyond the constraints of the original crime premise, such as detective-specific reasoning patterns, clue discovery sequences, and investigative behaviors. While this enriches the overall narrative, it may lead to story developments that partially deviate or extend beyond literal premise adherence, resulting in lower scores on a criterion specifically measuring premise matching. This trade-off suggests that the trait-guided method prioritizes narrative depth and investigative authenticity over strict premise constraints.

The engagement criterion presented a more complex pattern. While two models (Claude Opus 4.1 and Gemini 2.5 Pro) favored baseline stories and one (GPT-5) showed no significant difference, the lack of a significant interaction between generation approach (trait-guided vs. baseline) and detective ($F(6, 1246) = 0.48$, $p = .82$) indicates that this variable performance was not attributable to specific detective characteristics but rather reflected broader differences in how models interpret narrative engagement. The exceptionally low inter-rater agreement on engagement (ICC = 0.03), the lowest among all evaluated criteria, further highlights the subjective nature of this dimension. This finding is particularly notable given that Chhun et al. [41] reported moderate inter-rater agreement among human evaluators for engagement (ICC = 0.46), suggesting that while humans can achieve reasonable consensus on what makes a story engaging, this dimension poses particular challenges for consistent LLM-based evaluation. The human validation reinforces this concern: human evaluators showed a significant trait-guided preference on engagement, suggesting that the variability across LLM evaluators may obscure a benefit visible to human readers. This variability may result from the multifaceted nature of engagement itself, which encompasses elements such as narrative pacing, tension building, and individual reader preferences, all aspects that may be weighted differently by different models.

## 6. Concluding remarks

This study investigated whether large language models can generate detective stories that authentically reflect the investigative styles of classic fictional detectives while maintaining narrative quality. Building upon our preliminary work characterizing the investigative methods of seven iconic detectives through a multi-LLM analysis [13], we developed a trait-guided generation method that uses detective-specific investigative profiles to guide narrative composition. The method integrates these profiles with narrative principles grounded in literary

theory, particularly Todorov's dual narrative structure, to produce stories in which the selected detectives (Hercule Poirot, Sherlock Holmes, William Murdoch, Columbo, Father Brown, Miss Marple, and Auguste Dupin) solve crimes according to their characteristic investigative approaches. Through a two-phase evaluation study encompassing both detective style recognition and narrative quality assessment, we examined whether this trait-guided approach, as implemented in our fully operational prototype, successfully produces distinguishable, high-quality detective narratives across different crime scenarios and investigative styles.

Phase 1 of our evaluation demonstrated that the trait-guided method successfully produced detective-specific narratives with recognizable investigative characteristics. Using a multi-LLM identification approach across 22 models, we found that state-of-the-art LLMs achieved high accuracy in identifying detectives from anonymized stories based on their investigative styles, with the best-performing model (Claude Opus 4.1) reaching 82.4% accuracy. The identification patterns revealed differences in recognizability across detectives: Father Brown, Columbo, and several others achieved high recognition rates based primarily on their distinctive investigative methods, while William Murdoch presented consistent identification challenges due to overlapping characteristics with other analytically-oriented detectives. Qualitative analysis of identification explanations revealed that 56.2% of identifications were based on investigative methods, with the remaining cases involving contextual narrative elements such as linguistic patterns and behavioral characteristics. When restricting analysis to identifications based solely on investigative methods, overall accuracy remained high at 78.8%, validating that the trait-guided generation method effectively embedded distinctive reasoning patterns and investigative approaches into the generated narratives. These findings demonstrate that detective-specific investigative styles not only guided the generation process but manifested as recognizable patterns in the resulting stories.

Phase 2 examined whether the trait-guided method maintains or enhances narrative quality compared to a prompt ablation baseline. LLM-based evaluation across six quality criteria revealed that trait-guided stories achieved higher scores on four dimensions: coherence, empathy, surprise, and complexity, with all three evaluator models showing consistent directional agreement on these criteria. Length-controlled analyses confirmed robust condition effects on relevance, coherence, empathy, and engagement after covariate adjustment, while the differences observed for surprise and complexity may partly reflect the greater narrative detail and length produced by the trait-guided approach rather than the specific influence of investigative trait profiles on these narrative dimensions. The most substantial improvement appeared in empathy, with trait-guided stories scoring significantly higher across all three models, demonstrating enhanced emotional characterization and character depth. Per-detective analysis revealed that this empathy improvement varied by investigative style, with Father Brown demonstrating a particularly large effect, followed by other psychologically-oriented detectives such as Miss Marple, Hercule Poirot, and Columbo. In contrast, trait-guided stories scored lower on relevance, reflecting an expected trade-off where the incorporation of investigative characteristics and narrative depth led to story developments extending beyond strict premise constraints. For engagement, baseline stories achieved higher scores according to two of the three evaluator models, though the variable results and low inter-rater agreement (ICC = 0.03) indicated high subjectivity for this criterion. A preliminary human validation conducted on a balanced subset of 42 stories supported the central empathy finding through significant agreement between human and LLM evaluators, while showing that human evaluators favored trait-guided stories more strongly on engagement and surprise than the LLM evaluators did. Despite lower scores on relevance and mixed engagement results, the findings of Phase 2 demonstrate that the trait-guided method enhances multiple dimensions of narrative quality, particularly character depth and emotional portrayal, while maintaining comparable overall performance.

These findings offer several contributions to the field of computational narratology and AI-assisted creative writing. First, they demonstrate that large language models can generate detective stories that authentically preserve the investigative styles of classic fictional characters, as evidenced by high identification accuracy based on reasoning patterns and behavioral characteristics embedded in the narratives. Second, they show that the trait-guided method provides a practical approach to character-driven narrative generation within detective fiction, demonstrating how formalized character trait profiles can guide LLMs to produce stylistically distinctive narratives. Third, per-detective quality analysis consistently indicates that trait guidance produces character-aligned effects, with psychologically-oriented detectives showing particularly strong improvements in emotional depth, confirming that the method successfully translates trait profiles into distinctive narrative characteristics. Fourth, our publicly available dataset of 420 detective stories across seven characters and three crime scenarios, along with evaluation data from multiple LLMs, provides resources for future research in narrative generation and evaluation. Finally, the development of AI-Sleuths as a functional prototype demonstrates the practical feasibility of trait-guided narrative generation as an interactive tool for creative storytelling in the detective fiction domain.

Several limitations of this study should be acknowledged. First, while the anonymization process in Phase 1 effectively removed explicit detective identifiers, it did not fully eliminate all contextual narrative elements such as linguistic patterns, physical characteristics, and associated characters. In particular, broader structural elements such as narrative perspective and the presence of companion characters may remain diagnostic of specific detectives even after name and gender removal. More comprehensive anonymization would have required extensive rewriting of story content, potentially altering the investigative characteristics we sought to evaluate. Recognizing this limitation during the study design, we conducted a detailed qualitative analysis of identification reasoning and calculated adjusted accuracy metrics restricted to identifications based solely on investigative methods, which demonstrated that detective styles remained recognizable even when excluding contextual cues (78.8% accuracy). Second, the evaluation of narrative quality in Phase 2 primarily relied on LLM-based assessment complemented by a preliminary human validation study rather than a comprehensive human evaluation. While this approach follows a validated methodology [42] and enabled systematic evaluation at scale, using LLMs to evaluate LLM-generated content introduces an inherent circularity that remains a methodological limitation. The use of multiple evaluator models from different providers, combined with inter-rater agreement analysis and grounding in Chhun et al.'s validated framework, provides partial mitigation of this concern, though it does not fully substitute for human judgment. The preliminary human validation provides an additional mitigation, supporting the central empathy finding through significant human-LLM agreement and indicating that the LLM-based evaluation may be conservative on engagement and surprise. However, the small scale of the validation limits the precision of its findings, and a fuller human evaluation study with a larger and more diverse evaluator pool remains necessary to fully address the limitations of LLM-based evaluation. Third, the choice of Qwen 3 32B as the generation model, while practical for prototype development due to its open-source nature and ability to run on consumer hardware, represents a trade-off between accessibility and generation capability. More advanced proprietary models with stronger performance on complex reasoning tasks may produce stories with enhanced investigative style authenticity and narrative quality. Lastly, it is important to note that the stories generated and evaluated in our study consist of 12 to 18 narrative events, representing relatively short detective narratives.

Future research can build upon this work in multiple ways. While the preliminary human validation conducted in this study provides an initial step toward validating the LLM-based evaluation, a fuller human evaluation study with a larger and more diverse evaluator pool would

enable deeper exploration of reader preferences for different investigative styles, particularly regarding which detective characteristics or story elements are most effective in enhancing story quality. Beyond general narrative quality, future work could also investigate detective-fiction-specific evaluation criteria such as plot solvability, fair play clue placement, and temporal consistency, potentially establishing a validated assessment framework that complements the general criteria employed in this study. While the current study generated stories of 12 to 18 events, future work could explore longer detective narratives with more complex plot structures to assess whether investigative style consistency and narrative quality can be maintained across extended and more elaborated plots. The trait-guided generation framework demonstrated effectiveness within detective fiction, but its potential application to a different choice of investigators or even, after suitable adaptation, to other fictional characters and narrative genres represents another promising research direction. Characters from different literary traditions, such as adventurers, scientists, or political figures, possess distinctive behavioral and reasoning patterns that could similarly guide narrative generation, potentially establishing that formalized character trait profiles enhance AI-assisted storytelling across diverse narrative contexts. Additionally, evaluating different LLMs for story generation, particularly more capable models that demonstrated strong analytical performance in Phase 1 of our study, could reveal whether enhanced reasoning capabilities yield further improvements in investigative style authenticity and narrative quality.

This work demonstrates that computational methods can meaningfully engage with literary character analysis to generate narratives that preserve the distinctive styles of iconic fictional figures. By integrating formalized character trait profiles with narrative generation, we contribute to the field of computational narratology, illustrating how computational approaches can respect and capture the nuanced characteristics that define literary traditions. As large language models continue to advance, their role as collaborative tools for creative exploration offers new possibilities for interactive storytelling that builds upon established narrative conventions while enabling personalized creative experiences.

## CRediT authorship contribution statement

**Edirlei Soares de Lima:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Margot M.E. Neggers:** Writing – review & editing, Methodology, Formal analysis. **Marco A. Casanova:** Writing – review & editing. **Bruno Feijó:** Writing – review & editing. **Antonio L. Furtado:** Writing – review & editing, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Prompts used to characterize the investigative methods of fictional detectives

### A.1. Phase 1: Description generation prompt

Task: Generate a concise and formal description of the distinguishing characteristics of the investigative method used by the fictional detective $\{D_{name}\}$.

Requirements:

- Base your description of the investigative method used on stories where $\{D_{name}\}$ is the protagonist or a principal investigator.
- Focus solely on the investigative approach, strategies, and distinguishing features that define this detective's method of solving cases.
- Consider only the distinguishing characteristics that set the investigative method of $\{D_{name}\}$ apart from those of other fictional detectives.
- Do not include biographical details, story summaries, or references to specific cases.
- Structure the response as a single paragraph, without bullet points or numbered lists.
- Do not include any introductory or concluding sentences outside of the description itself.
- Limit the response to a maximum of 5 sentences.

### A.2. Phase 2: Trait extraction prompt

Extract the key traits that describe the investigative method in the following text. List each trait as a separate bullet point. Use a formal and concise style. Do not repeat traits, and do not add information that is not present in the text.

*Text:* $\{D_{description}\}$

### A.3. Phase 3: Semantic grouping prompt

List of Traits:

$\{D_{traits}\}$

Task: Given the list of traits above describing the investigative methods of a fictional detective, group together all traits that express the same or highly similar idea, regardless of phrasing or wording.

For each group, provide:

- A description of the core idea of the investigative method, taking into account the grouped traits. Use all relevant distinguishing terms when writing the description.
- The list of original traits belonging to the group.

Present the output as a JSON array in the following format:

```
[
 {
  "label": "Description of the core idea of the investigative method using relevant distinguishing terms",
  "traits": ["Original phrasing 1", "Original phrasing 2"]
 }
]
```

### A.4. Phase 5: Validation via reverse identification prompt

List of Traits:

$\{D_{profile}\}$

You are given a list of traits describing the investigative method of a fictional detective. Your task is to identify which detective this description most likely refers to.

Choose only from the following list:

- Hercule Poirot
- Sherlock Holmes
- William Murdoch
- Columbo

- Father Brown
- Miss Marple
- Auguste Dupin

Respond only with the name of the detective, without explanations or additional text.

## Appendix B. List of investigative traits identified for each detective

| Detective | Trait description |
|---|---|
| Hercule Poirot | Reliance on psychological insight and understanding of human nature<br>Meticulous attention to detail and observation of human behavior and psychology<br>Systematic and methodical approach to gathering and analyzing information<br>Emphasis on logical deduction and reasoning<br>Reconstruction of events through logical inferences and mental synthesis<br>Use of conversation and interviews to extract information<br>Focus on motive and psychological underpinnings<br>Use of intuition and insight to identify inconsistencies and hidden clues<br>Emphasis on order, symmetry, and method<br>Use of theatrical flair and dramatic confrontations<br>Presentation of a comprehensive and internally consistent solution |
| Sherlock Holmes | Extraordinary ability to observe minute details and deduce complex conclusions<br>Application of logical reasoning and deduction to solve mysteries<br>Systematic and analytical approach to investigation<br>Utilization of scientific knowledge and principles in investigation<br>Extensive knowledge across various disciplines<br>Reconstruction of events through physical clues and mental simulation<br>High degree of experimentation and verification<br>Psychological profiling and understanding of human behavior<br>Use of forensic science and empirical evidence<br>Use of creative thinking and unconventional methods<br>Use of disguises and undercover strategies for information gathering<br>Synthesis of disparate facts into coherent narratives<br>Elimination of impossibilities to determine the truth |
| William Murdoch | Meticulous and scientific approach to evidence collection and analysis<br>Use of forensic science and innovative techniques<br>Application of experimental methods and technologies<br>Rigorous observation, deductive reasoning, and logical analysis<br>Use of early forensic techniques such as fingerprinting and blood analysis<br>Reconstruction of crimes and crime scenes<br>Understanding of human psychology and behavior<br>Empirical observation and evidence-based reasoning<br>Openness to new ideas and challenging conventional wisdom<br>Identification of subtle patterns and connections<br>Collaboration with experts from various fields<br>Use of photography and documentation techniques |
| Columbo | Use of psychological manipulation and understanding of human psychology<br>Noting and piecing together insignificant facts and inconsistencies<br>Disarming and unassuming demeanor to lull suspects into a false sense of security<br>Uncovering contradictions and revealing hidden truths through questioning<br>Meticulous attention to detail and acute observational skills<br>Combination of deceptive demeanor with rigorous observation and analytical skills<br>Feigning forgetfulness or confusion as a strategic tool to manipulate interactions<br>Revisiting and reexamining evidence and crime scenes<br>Strategy of persistent questioning and returning to suspects with additional queries<br>Building rapport with witnesses and suspects through casual conversation<br>Use of intuition and patience over rapid breakthroughs<br>Meticulous reconstruction of the crime scene through incremental evidence gathering<br>Securing confessions through the suspect's own words |
| Father Brown | Profound intuition and deep understanding of human nature<br>Emphasis on intuitive insight and psychological understanding<br>Observation of subtleties of human behavior and nuances of moral character<br>Uncovering motives and intentions that elude conventional detectives<br>Ability to identify and challenge preconceptions<br>Combination of theology and philosophy with practical knowledge of human frailty<br>Ability to blend into various social settings<br>Focus on the why of a crime rather than the how<br>Humble and unassuming demeanor<br>Anticipating actions and thoughts of criminals<br>Solving crimes by understanding the type of sinner rather than assembling physical proof<br>Distinction from empirically focused detective methods through psychological acuity and moral understanding<br>Distinctive blend of psychological acuity, moral awareness, and intellectual curiosity<br>Subtle grasp of moral and spiritual insights<br>Drawing out confessions or revelations through conversation<br>Presumption of inherent goodness in individuals<br>Empathetic reconstruction of the criminal's moral fall<br>Identification of perpetrators through recognition of concealed regret or desperation<br>Use of humility and empathy<br>Emphasis on understanding motivations and emotional states of individuals<br>Focuses on discerning motives and moral weaknesses |

| | |
|---|---|
| Miss Marple | Reliance on keen observation and acute understanding of human nature and behavior<br>Use of psychological insights and understanding of social dynamics<br>Utilization of extensive knowledge of village life and its inhabitants<br>Gathering information through indirect methods such as conversations and social interactions<br>Making connections between past experiences and current events<br>Construction of a comprehensive picture of circumstances through observation, deduction, and subtle manipulation of conversations<br>Assessment of individuals' credibility and motives without direct questioning or physical evidence<br>Use of age and apparent innocence to disarm suspects and gather information<br>Listening attentively to gossip and trivial details to uncover hidden connections<br>Emphasis on patience, empathy, and intuitive understanding of human frailties<br>Leveraging memory and understanding of human nature to identify patterns and motives |
| Auguste Dupin | Understanding and anticipation of human psychology, behavior, and motivations<br>Rigorous and analytical reasoning through observation, deduction, and logical analysis (ratiocination)<br>Meticulous observation and attention to detail, noticing subtle cues<br>Use of imagination and creative thinking to explore unconventional solutions<br>Prioritization of mental processes and intellectual analysis over physical evidence<br>Reconstruction of events and uncovering hidden evidence through logical deduction<br>Recognition of solutions concealed in plain sight and challenge of conventional assumptions |

## Appendix C. Story generation prompt

Protagonist: $\{D_{name}\}$.
Characteristics of the Investigative Method: $\{D_{methods}\}$.
Crime Premise: $\{C_{premise}\}$.

Your task is to write a detective story in a structured format, based on the given crime premise, protagonist, and set of investigative characteristics that define the detective's investigative style. The protagonist must solve the case by acting in accordance with the listed investigative method, but you must not state these characteristics explicitly in the story. Instead, they should be demonstrated naturally through the detective's observations, actions, deductions, dialogue, and interactions.

Use Tzvetan Todorov's dual narrative structure as an inspiration:

- Reveal the story of the crime gradually, through the detective's investigation.
- Let the investigation be the visible driving narrative.

The story should include:

- A crime or apparent crime, which can turn out to be a deception, accident, or collective act.
- A series of clues, inconsistencies, or observations that others overlook.
- The detective's reasoning, often abductive, should turn these clues into hypotheses.
- Serendipitous or seemingly random details should become meaningful through the detective's insight.
- A final reconstruction of what actually happened must be presented at the end, grounded in the investigation.
- The solution must be convincing and feel like a natural consequence of the discovered evidence.
- Subtle aspects of the detective's personality, even if unrelated to the case.
- Natural-sounding dialogue when characters interact. Use dialogue to: expose hidden motivations, mislead or redirect attention, allow the detective to probe, challenge, or reflect, and to build tension or highlight personal traits.

Format Instructions:

- First, generate a creative title for the story on a line starting with: *"TITLE:"*
- Then, generate the entire story as a sequence of narrative events. Each event should follow this format:
  - *"EVENT X:"* followed by a scene description, including dialogue when needed (X is the event number, starting from 1).
  - *"IMAGE X:"* followed by a short description of an image that could illustrate that event. Use the full character names when mentioning them in the image description.
- Number each event sequentially (e.g., EVENT 1, IMAGE 1, EVENT 2, IMAGE 2, ...) to clearly mark the story progression.
- Format the event descriptions using Markdown to improve readability, with a new paragraph for each character's dialogue or shift in action within the event.
- The story should consist of approximately 12 to 18 events, covering the full arc from the introduction to the final resolution.

## Appendix D. Story regeneration prompt

Protagonist: $\{D_{name}\}$.
Characteristics of the Investigative Method: $\{D_{methods}\}$.
Crime Premise: $\{C_{premise}\}$.
Full Story:
$\{S_{title}\}$
$\{S_{events}\}$

The user has suggested a change to Event $\{E_{index}\}$: $\{S_{suggestion}\}$

Your task is to regenerate Event $\{E_{index}\}$ based on this suggestion, while maintaining the story's coherence, tone, and logic. You may rewrite later events (from Event $\{E_{index}\}$ onward) only if necessary to ensure consistency with the updated event.

Use Tzvetan Todorov's dual narrative structure as an inspiration:

- Reveal the story of the crime gradually, through the detective's investigation.
- Let the investigation be the visible driving narrative.

The story should include:

- A crime or apparent crime, which can turn out to be a deception, accident, or collective act.
- A series of clues, inconsistencies, or observations that others overlook.
- The detective's reasoning, often abductive, should turn these clues into hypotheses.
- Serendipitous or seemingly random details should become meaningful through the detective's insight.
- A final reconstruction of what actually happened must be presented at the end, grounded in the investigation.
- The solution must be convincing and feel like a natural consequence of the discovered evidence.

- Subtle aspects of the detective's personality, even if unrelated to the case.
- Natural-sounding dialogue when characters interact. Use dialogue to: expose hidden motivations, mislead or redirect attention, allow the detective to probe, challenge, or reflect, and to build tension or highlight personal traits.

Format Instructions:

- Regenerate the story starting from Event $\{E_{index}\}$. Do not change, repeat, or summarize any events before Event $\{E_{index}\}$.
- Event numbering must begin at $\{E_{index}\}$ and continue sequentially (e.g., EVENT $\{E_{index}\}$, IMAGE $\{E_{index}\}$, EVENT $\{E_{index}+1\}$, IMAGE $\{E_{index}+1\}$, ...).
- Each event should follow this format:
  - *"EVENT X:"* followed by a scene description, including dialogue when needed.
  - *"IMAGE X:"* followed by a short description of an image that could illustrate that event. Use the full character names when mentioning them in the image description.
- Format the event descriptions using Markdown to improve readability, with a new paragraph for each character's dialogue or shift in action within the event.
- The regenerated events must preserve narrative consistency with the earlier events, while addressing the user's suggestion.
- The full story (including unchanged earlier events) should still consist of approximately 12 to 14 events in total.

## Appendix E. Crime premise enhancer prompt

Original Crime Premise: $\{C_{premise}\}$.
Characteristics of the Investigative Method: $\{D_{methods}\}$.
Your task is to improve the provided crime premise to make it more detailed and better suited for an engaging detective story. The goal is to keep the original intent and situation intact, but enhance it so that:

- It includes enough details and complexity to allow a meaningful investigation.
- It creates room for deductive or abductive reasoning, with potential for clues, suspects, false leads, or ambiguous circumstances.

Do not change the core elements described in the crime premise. Instead, add depth, ambiguity, and opportunity for logical or psychological analysis. Do not add any information about the detective.
Write the improved crime premise as a single and concise paragraph limited to a maximum of 3 sentences.

## Appendix F. Baseline story generation prompt

Write a detective story where $\{D_{name}\}$ investigates and solves the following crime:
$\{C_{premise}\}$
Format Instructions:

- First, generate a creative title for the story on a line starting with: *"TITLE:"*
- Then, generate the entire story as a sequence of narrative events. Each event should follow this format:
  - *"EVENT X:"* followed by a scene description, including dialogue when needed (X is the event number, starting from 1).
  - *"IMAGE X:"* followed by a short description of an image that could illustrate that event. Use the full character names when mentioning them in the image description.
- Number each event sequentially (e.g., EVENT 1, IMAGE 1, EVENT 2, IMAGE 2, ...) to clearly mark the story progression.
- Format the event descriptions using Markdown to improve readability, with a new paragraph for each character's dialogue or shift in action within the event.
- The story should consist of approximately 12 to 18 events, covering the full arc from the introduction to the final resolution.

## Appendix G. Story evaluation framework

### G.1. Story evaluation prompt

Premise: $\{C_{premise}\}$.
Story:
$\{S_{title}\}$
$\{S_{events}\}$
Task: Critically evaluate and rate the provided story on $\{E_{criterion}\}$ using the scale presented below.
$\{E_{criterion}\}$: $\{E_{description}\}$
Rating Scale:
$\{E_{guidelines}\}$
Output format:
RATE: [number from 1–5]
EXPLANATION: [2–3 sentences citing specific story elements that justify your rating]

### G.2. Story evaluation criteria and guidelines

See Table 9.

**Table 9**
Story evaluation criteria and guidelines.

| Criterion | Description | Guidelines |
|---|---|---|
| Relevance | How well the story addresses and develops its premise. | 1: The story has no discernible connection to the premise.<br>2: The story has only a weak or tangential relationship with the premise.<br>3: The story addresses the main elements of the premise but with notable gaps or deviations.<br>4: The story matches the premise well, with only one or two minor aspects not fully addressed.<br>5: The story matches the premise completely, developing all its key elements. |
| Coherence | Whether the story maintains internal consistency and logical flow. | 1: The story does not make sense. The setting and/or characters are contradictory, and/or there is no understandable plot.<br>2: Most of the story contains logical inconsistencies or narrative confusion.<br>3: The story is mostly coherent but contains some noticeable inconsistencies in logic, character behavior, or plot.<br>4: The story is coherent overall, with only one or two minor inconsistencies that do not significantly impact understanding.<br>5: The story maintains complete logical consistency from beginning to end. |

(*continued on next page*)

**Table 9** (*continued*).

| Criterion | Description | Guidelines |
|---|---|---|
| Empathy | How clearly and effectively the story portrays characters' emotional states. | 1: The characters display no emotional depth or relatable feelings.<br>2: At least one character shows basic emotions that create minimal emotional connection.<br>3: At least one character displays specific, recognizable emotions (e.g., sadness, joy, fear) in a believable way.<br>4: At least one character creates strong emotional resonance, though minor elements limit full relatability.<br>5: At least one character demonstrates deep emotional complexity that feels authentic and compelling. |
| Surprise | How unpredictable the story's ending is relative to the clues provided. | 1: The ending is either completely obvious from the start, or contradicts the story without making sense.<br>2: The ending becomes easily predictable within the first few paragraphs.<br>3: The ending becomes predictable around the story's midpoint.<br>4: The ending is surprising and would have been difficult to predict beforehand.<br>5: The ending is both surprising and well-foreshadowed. It feels unexpected yet inevitable in retrospect. |
| Engagement | How effectively the story maintains narrative momentum and reader interest. | 1: The story lacks compelling elements and fails to maintain interest.<br>2: The story has one or two mildly interesting elements but is largely unengaging.<br>3: The story maintains moderate interest with some engaging moments.<br>4: The story is engaging throughout with only brief less compelling passages.<br>5: The story is highly engaging with strong narrative pull that sustains interest completely. |
| Complexity | How elaborate and well crafted the story's elements are (plot, characters, setting, themes). | 1: The story presents events without motivation, consequence, or connection.<br>2: Characters act predictably without depth, events follow obvious patterns, or plot points feel arbitrary.<br>3: Characters have clear motivations, plot follows logical cause-and-effect, and details serve functional purposes.<br>4: Characters reveal psychological depth through actions, plot emerges organically from character choices, and details enhance atmosphere or meaning.<br>5: Characters embody contradictions or evolve authentically, narrative structure creates interpretive layers, and elements interconnect to generate resonance. |

## Data availability

Data will be made available on request.

## References


[1] M.-L. Ryan, Interactive narrative, plot types, and interpersonal relations, in: U. Spierling, N. Szilas (Eds.), Interactive Storytelling, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 6–13, http://dx.doi.org/10.1007/978-3-540-89454-4_2.

[2] Aristotle, in: M. Heath (Ed.), Poetics, Penguin Classics, London, 1995, 1995, originally written c. 335 BCE.

[3] H. Walpole, Letter to Horace Mann, January 28, 1754, Letter, 1754, URL: https://archive.org/details/lettersofhoracew01walp/page/408.

[4] C.S. Peirce, in: N. Houser, C. Kloesel (Eds.), The Essential Peirce: Selected Philosophical Writings, Volume 1 (1867–1893), Indiana University Press, Bloomington, 1992, 1992.

[5] U. Eco, T.A. Sebeok (Eds.), The Sign of the Three: Dupin, Holmes, Peirce, Indiana University Press, Bloomington, 1988.

[6] T. Todorov, The typology of detective fiction, in: T. Todorov (Ed.), The Poetics of Prose, Cornell University Press, Ithaca, 1977, pp. 42–52.

[7] S.D. Barbosa, E.S. de Lima, A.L. Furtado, B. Feijó, Generation and dramatization of detective stories, J. Interact. Syst. 5 (2) (2014) http://dx.doi.org/10.5753/jis.2014.648.

[8] C. Jaschek, T. Beckmann, J.A. Garcia, W.L. Raffe, Mysterious murder - MCTS-driven murder mystery generation, in: 2019 IEEE Conference on Games (CoG), 2019, pp. 1–8, http://dx.doi.org/10.1109/CIG.2019.8848000.

[9] Y.-G. Cheong, R.M. Young, Suspenser: A story generation system for suspense, IEEE Trans. Comput. Intell. AI Games 7 (1) (2015) 39–52, http://dx.doi.org/10.1109/TCIAIG.2014.2323894.

[10] R. Doust, P. Piwek, A model of suspense for narrative generation, in: J.M. Alonso, A. Bugarín, E. Reiter (Eds.), Proceedings of the 10th International Conference on Natural Language Generation, Association for Computational Linguistics, Santiago de Compostela, Spain, 2017, pp. 178–187, http://dx.doi.org/10.18653/v1/W17-3527.

[11] V. Kumaran, J. Rowe, B. Mott, J. Lester, SceneCraft: Automating interactive narrative scene generation in digital games with large language models, Proc. AAAI Conf. Artif. Intell. Interact. Digit. Entertain. 19 (1) (2023) 86–96, http://dx.doi.org/10.1609/aiide.v19i1.27504.

[12] V. Kumaran, J. Rowe, J. Lester, NarrativeGenie: Generating narrative beats and dynamic storytelling with large language models, Proc. AAAI Conf. Artif. Intell. Interact. Digit. Entertain. 20 (1) (2024) 76–86, http://dx.doi.org/10.1609/aiide.v20i1.31868.

[13] E.S. de Lima, M.A. Casanova, B. Feijó, A.L. Furtado, Characterizing the investigative methods of fictional detectives with large language models, 2025, URL: https://arxiv.org/abs/2505.07601, arXiv:2505.07601.

[14] H. Mohr, M. Eger, C. Martens, Eliminating the impossible: A procedurally generated murder mystery, in: J. Zhu (Ed.), Joint Proceedings of the AIIDE 2018 Workshops, Co-Located with the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2018), in: CEUR Workshop Proceedings, vol. 2282, Edmonton, Canada, 2018, pp. 1–7, URL: https://ceur-ws.org/Vol-2282/EXAG_113.pdf.

[15] G.A. Barros, A. Liapis, J. Togelius, Murder mystery generation from open data, in: F. Pachet, F.A. Cardoso, V. Corruble, F. Ghedini (Eds.), Proceedings of the 7th International Conference on Computational Creativity, ICCC 2016, Paris, France, 2016, pp. 197–204.

[16] R. Zhao, Q. Zhu, H. Xu, J. Li, Y. Zhou, Y. He, L. Gui, Large language models fall short: Understanding complex relationships in detective narratives, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7618–7638, http://dx.doi.org/10.18653/v1/2024.findings-acl.454.

[17] Z. Xu, J. Ye, X. Liu, X. Liu, T. Sun, Z. Liu, Q. Guo, L. Li, Q. Liu, X. Huang, X. Qiu, DetectiveQA: Evaluating long-context reasoning on detective novels, 2024, http://dx.doi.org/10.48550/arXiv.2409.02465, arXiv:2409.02465.

[18] E. Wagner, R. Keydar, O. Abend, Modeling fair play in detective stories with language models, 2025, http://dx.doi.org/10.48550/arXiv.2507.13841, arXiv:2507.13841.

[19] K. Xie, M. Riedl, Creating suspenseful stories: Iterative planning with large language models, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Malta, 2024, pp. 2391–2407, http://dx.doi.org/10.18653/v1/2024.eacl-long.147.

[20] D. Iudin, Procedural Generator of Short Detective-Like Stories (Master' thesis), Charles University, Faculty of Mathematics and Physics, Department of Software and Informatics Education, Prague, 2023.

[21] A. Fan, M. Lewis, Y. Dauphin, Hierarchical neural story generation, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 889–898, http://dx.doi.org/10.18653/v1/P18-1082.

[22] S. Värtinen, P. Hämäläinen, C. Guckelsberger, Generating role-playing game quests with GPT language models, IEEE Trans. Games 16 (1) (2024) 127–139, http://dx.doi.org/10.1109/TG.2022.3228480.

[23] A. Alabdulkarim, W. Li, L.J. Martin, M.O. Riedl, Goal-directed story generation: Augmenting generative language models with reinforcement learning, 2021, http://dx.doi.org/10.48550/arXiv.2112.08593, arXiv:2112.08593.

[24] P. Xu, M. Patwary, M. Shoeybi, R. Puri, P. Fung, A. Anandkumar, B. Catanzaro, MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 2831–2845, http://dx.doi.org/10.18653/v1/2020.emnlp-main.226.

[25] L. Castricato, S. Frazier, J. Balloch, M. Riedl, Tell me a story like i'm five: Story generation via question answering, in: Proceedings of the 3rd Workshop on Narrative Understanding, 2021, pp. 1–8, URL: https://par.nsf.gov/biblio/10249509.

[26] K. Yang, Y. Tian, N. Peng, D. Klein, Re3: Generating longer stories with recursive reprompting and revision, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4393–4479, http://dx.doi.org/10.18653/v1/2022.emnlp-main.296.

[27] A. Gurung, M. Lapata, Learning to reason for long-form story generation, 2025, http://dx.doi.org/10.48550/arXiv.2503.22828, arXiv:2503.22828.

[28] J. Li, Y. Chen, Z. Liu, M. Tan, L. Zhang, Y. Li, R. Luo, L. Chen, J. Luo, A. Argha, H. Alinejad-Rokny, W. Zhou, M. Yang, STORYTELLER: An enhanced plot-planning framework for coherent and cohesive story generation, in: W. Che, J. Nabende, E. Shutova, M.T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 20818–20846, http://dx.doi.org/10.18653/v1/2025.findings-acl.1071.

[29] S. Venkatraman, N.I. Tripto, D. Lee, CollabStory: Multi-LLM collaborative story generation and authorship analysis, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 3665–3679, http://dx.doi.org/10.18653/v1/2025.findings-naacl.203.

[30] E.S. de Lima, M.M.E. Neggers, M.A. Casanova, B. Feijó, A.L. Furtado, A pattern-oriented AI-powered approach to story composition, in: P. Figueroa, A. Di Iorio, D. Guzman del Rio, E.W. Gonzalez Clua, L. Cuevas Rodriguez (Eds.), Entertainment Computing – ICEC 2024, Springer Cham, 2024, pp. 1–16, http://dx.doi.org/10.1007/978-3-031-74353-5_10.

[31] E.S. de Lima, M.A. Casanova, A.L. Furtado, Imagining from images with an AI storytelling tool, 2024, URL: https://arxiv.org/abs/2408.11517. arXiv:2408.11517.

[32] E.S. de Lima, M.M. Neggers, M.A. Casanova, A.L. Furtado, From images to stories: Exploring player-driven narratives in games, in: A. Marto, R. Prada, P. Gouveia, R. Contreras-Espinosa, A. Gon¸calves, E. Abrantes, R. Ribeiro (Eds.), Videogame Sciences and Arts, Springer Nature Switzerland, Cham, 2025, pp. 228–242, http://dx.doi.org/10.1007/978-3-031-81713-7_16.

[33] E.S. de Lima, B. Feijó, A.L. Furtado, Computational narrative blending based on planning, in: J. Baalsrud Hauge, J. C. S. Cardoso, L. Roque, P.A. Gonzalez-Calero (Eds.), Entertainment Computing – ICEC 2021, Springer International Publishing, Cham, 2021, pp. 89–303, http://dx.doi.org/10.1007/978-3-030-89394-1_22.

[34] E.S. de Lima, B. Feijó, M.A. Casanova, A.L. Furtado, ChatGeppetto - an AI-powered storyteller, in: Proceedings of the 22nd Brazilian Symposium on Games and Digital Entertainment, ACM, 2024, pp. 28–37, http://dx.doi.org/10.1145/3631085.3631302.

[35] E.S. de Lima, M.A. Casanova, B. Feijó, A.L. Furtado, Semiotic structuring in movie narrative generation, in: P. Ciancarini, A. Di Iorio, H. Hlavacs, F. Poggi (Eds.), Entertainment Computing – ICEC 2023, Springer Nature Singapore, Singapore, 2023, pp. 161–175, http://dx.doi.org/10.1007/978-981-99-8248-6_13.

[36] E.S. de Lima, M.M. Neggers, B. Feijó, M.A. Casanova, A.L. Furtado, An AI-powered approach to the semiotic reconstruction of narratives, Entertain. Comput. 52 (2025) 100810, http://dx.doi.org/10.1016/j.entcom.2024.100810.

[37] J.G. Cawelti, Adventure, Mystery, and Romance: Formula Stories as Art and Popular Culture, University of Chicago Press, Chicago, 1976.

[38] J. Scaggs, Crime Fiction, Routledge, London and New York, 2005.

[39] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025, http://dx.doi.org/10.48550/arXiv.2505.09388, arXiv:2505.09388.

[40] E. van Laethem, Connecting Detectives: Investigation of the Extent to Which Arthur Conan Doyle Was Influenced by Edgar Allan Poe's Detective Auguste Dupin When Creating Sherlock Holmes (Master' thesis), Ghent University, Faculty of Arts and Philosophy, 2017.

[41] C. Chhun, P. Colombo, F.M. Suchanek, C. Clavel, Of human criteria and automatic metrics: A benchmark of the evaluation of story generation, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T.K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 5794–5836, URL: https://aclanthology.org/2022.coling-1.509/.

[42] C. Chhun, F.M. Suchanek, C. Clavel, Do language models enjoy their own stories? prompting large language models for automatic, Trans. Assoc. Comput. Linguist. 12 (2024) 1122–1142, http://dx.doi.org/10.1162/tacl_a_00689.