WhoDunIt: Evaluation benchmark for culprit detection in mystery stories

Anonymous ACL submission

Abstract

We present a novel data set, WhoDunIt, to assess the deductive reasoning capabilities of large language models (LLM) within narrative contexts. Constructed from open domain mystery novels and short stories, the dataset challenges LLMs to identify the perpetrator after reading and comprehending the story. To evaluate model robustness, we apply a range of character-level name augmentations, including original names, name swaps, and substitutions with well-known real and/or fictional entities from popular discourse. We further use various prompting styles to investigate the influence of prompting on deductive reasoning accuracy.

We conduct evaluation study with state-of-the-art models, specifically *GPT-4o, GPT-4-turbo,* and *GPT-4o-mini*, evaluated through multiple trials with majority response selection to ensure reliability. The results demonstrate that while LLMs perform reliably on unaltered texts, accuracy diminishes with certain name substitutions, particularly those with wide recognition.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional capabilities in a wide array of natural language tasks, from text generation and summarization to complex reasoning and inference (Brown, 2020). The release of the transformer architecture by Vaswani (2017) marked a pivotal advancement in the field, enabling models to handle long-range dependencies in text more effectively through self-attention mechanisms. This breakthrough not only enhanced model scalability but also laid the foundation for the development of increasingly sophisticated LLMs that are now capable of handling nuanced and context-rich tasks. With the emergence of models such as BERT(Kenton and Toutanova, 2019), GPT-2(Radford et al., 2019), and later ChatGPT(OpenAI, 2022), the field of natural language processing has seen rapid innovation, driving significant improvements in model performance and expanding potential applications.

ChatGPT demonstrated that LLMs could deliver highly interactive, contextually relevant responses in real-time, broadening their accessibility to non-technical users and sparking widespread integration in industries. This release emphasized the need for systematic evaluation frameworks to understand the capabilities, limitations, and potential biases of these models as they are adopted in real-world applications.

Over recent years, several significant benchmarks have been introduced, such as MMLU(Hendrycks et al., 2020), HELM(Liang et al., 2022), Open LLM Leaderboard[1], and AlpacaEval[2]. These benchmarks have been critical in capturing LLM reasoning capabilities and enabling comparisons among state-of-the-art models.

This paper contributes to these efforts by introducing a novel dataset specifically designed to assess deductive reasoning within narrative contexts. To build this dataset we take inspiration from a recent interview(Huang and Sutskever, 2023) between Ilya Sutskever and Jensen Huang about "next word prediction" being sufficient for understanding. Our benchmark aims to provide deeper insights into the adaptability and inference capabilities of leading models, including *GPT-4o, GPT-4-turbo,* and *GPT-4o-mini* (Achiam et al., 2023), especially in tasks involving complex narrative comprehension. We believe that such a benchmark will help future model iteration on LLMs deductive reasoning capabilities as well as complex long-form narrative comprehension.

This paper is organized as follows: Section 2 reviews relevant prior research, while Section 3 de-

[1] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

[2] https://github.com/tatsu-lab/alpaca_eval

tails the dataset preparation process. In Section 4, we describe the experimental setup used for evaluation. Section 5 presents our findings and analyzes them in terms of LLM capabilities. Finally, Section 6 offers conclusions and outlines directions for future work.

2 Related Works

Foundational LLMs, such as GPT-2 and GPT-3, demonstrated strong performance across various text-based tasks, though they initially struggled with complex, multi-step reasoning (Radford et al., 2019; Brown, 2020).

CoT prompting, which encourages models to break down problems into logical steps, has been shown to enhance accuracy and coherence in deductive tasks (Wei et al., 2022). Additional methods, like Self-Reflection prompting, further improve reliability by having models verify and refine their responses, leading to more thoughtful answers (Shinn et al., 2024; Madaan et al., 2024).

LLMs' abilities to handle narrative reasoning—tracking characters, plot progression, and thematic elements—have also been a focal area of AI research. Studies have shown that while models can generate coherent stories, they often struggle with consistency over long narratives (Ammanabrolu et al., 2021; Rashkin et al., 2020). Enhanced approaches have aimed to improve narrative coherence, though challenges remain, particularly in maintaining character roles and logical plot flow.

Several benchmarks assess LLMs' reasoning and comprehension, including MMLU, HELM, and Big-Bench (BBH), which evaluate performance across diverse tasks (Hendrycks et al., 2020; Liang et al., 2022; Srivastava et al., 2022). These benchmarks incorporate tasks requiring reasoning and narrative comprehension, though few focus specifically on deductive reasoning within mystery narratives.

3 Dataset Preparation

In this section, we outline our dataset preparation, validation process. To release this dataset for open source use, we focus on books that have entered the public domain, so we use *Project Gutenberg*[3] as our primary story source. We then obtained the list of 500+ Mystery and Detective story titles, that are of interest to us. Additionally, to maintain sufficient variability and diversity in the dataset, we ensured that we represent all the broad characteristics of the stories. Each selected novel features an identifiable culprit, ensuring that the task involves pinpointing to perpetrator. The novels span a diverse range of authors and storytelling styles, encompassing classic *WhoDunIt* detective novels by authors such as Agatha Christie. As shown in Figure 1, the stories vary in length, covering short, medium and full narratives, providing a broad spectrum of text. By including works from different writers and narrative traditions, we ensure that the models encounter a variety of narrative structures, reasoning styles, and linguistic expressions used to describe mystery and crime.

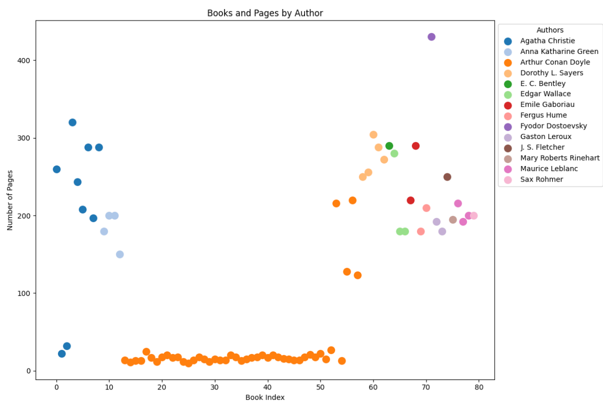

Figure 1: Distribution by Length

Since these stories are very popular and have been in the public discourse for a long time, for most of the stories, we find the identity of the culprit from services like *Cliffnotes*[4]. This provides us confidence about the identity of the culprit of the story, and hence the accuracy of our dataset. Secondly, for others we read them ourselves to figure out the culprit of the story.

Since these stories are in public domain, any model has most likely already been trained on them. Additionally, model would also have trained on any notes/blog posts about these stories. Thus the identity of culprit is probably already in model's memory. To further investigate whether the model depends on memorized data from pre-training or can genuinely engage in contextual reasoning, we applied a series of character-name substitutions. Each augmentation is intended to disrupt potential memorized associations with names, forcing the model to rely on contextual cues and relationships between characters, rather than merely recognizing famous names.

[3] https://www.gutenberg.org/

[4] https://www.cliffsnotes.com/

Here are the specific augmentations and the rationale behind each:

- **Original Character Names:** This serves as a control, where no modifications are made to the text, providing a baseline for the model's deduction capabilities with familiar, unchanged names.
- **Full Character Name Swap:** Here, we swap the names of all characters in the story. This approach is intended to test the model's capacity to follow complex character interactions and relationships without relying on the original names. This alteration simulates a scenario where familiar identifiers are altered, requiring the model to deduce based on narrative function rather than name recognition.
- **Replacement with Harry Potter Character Names:** In this augmentation, we replace all character names with those of well-known characters from the Harry Potter series. This tactic tests the model's ability to ignore pre-trained associations tied to widely recognized fictional characters, focusing instead on the plot's internal logic and character roles within the story.
- **Hollywood Celebrity Names:** Replacing names with those of famous Hollywood celebrities introduces a real-world layer of familiarity, which can potentially interfere with the model's reasoning if it relies on pre-trained biases. This approach assesses the model's ability to disregard prominent, real-world associations and concentrate solely on the characters' roles within the narrative structure.
- **Bollywood Celebrity Names:** Similarly, substituting names with Bollywood celebrities introduces an additional layer of cultural recognition. This augmentation not only adds diversity to the test but also evaluates whether the model can apply the same deductive process across different cultural references, further examining its adaptability and robustness under diverse, globally recognizable identities.

By applying these augmentation techniques, we systematically modify the dataset to create various degrees of reasoning difficulty, thus challenging the LLM's deductive capabilities in unique ways. Each augmentation serves to disrupt familiar name associations, encouraging the model to prioritize contextual understanding and narrative roles over memorized patterns or recognizable identities.

The list of novels used can be found in the Appendix A.3 and few examples of the point of reveal in stories of our dataset[5] can be found in the Appendix A.1, A.2.

4 Experimental Setup

We conducted our experiments on three OpenAI models: *GPT-4o, GPT-4-turbo,* and *GPT-4o-mini* (Achiam et al., 2023), using OpenAI's Batch API[6] via the chat-completions endpoint. These models represent a spectrum of capabilities within the GPT-4 family, allowing us to examine how model size and design impact performance in narrative deduction tasks.

4.1 Prompting Techniques

To assess the models' reasoning abilities, we applied four prompting styles:

1. **Basic Prompting**: Basic prompting without additional guidance, providing a baseline for model performance (Brown, 2020).
2. **Self-Reflection Prompting**: The model is encouraged to review its response for accuracy, simulating a reflective process that can improve answer quality (Shinn et al., 2024).
3. **Chain-of-Thought(CoT) Prompting**: Instructs the model to reason through tasks step-by-step, enhancing clarity and accuracy in complex problem-solving (Wei et al., 2022).
4. **CoT + Self-Reflection**: Combines step-by-step reasoning with self-reflection, prompting the model to refine its answer after an initial response for improved reliability (Madaan et al., 2024).

To reduce the variability of responses, and ensure we capture the maximum level of LLM reasoning, we consider a 10-shot prompting for each prompt variety and use the most frequent response as the answer(Wang et al., 2022).

With basic prompt as baseline, the self-reflexion is better than that signifying that reflective check fairly improves the performance and adding COT

[5]It will be public at the time of submission to a conference

[6]https://platform.openai.com/docs/guides/batch/overview

to both of these add fairly to the accuracy of the system.

5 Results and Analysis

To ensure robust and reliable results, we evaluated each model's performance by conducting 10 independent calls for each configuration(Wang et al., 2022). In each trial, we maintained consistent input conditions—specifically, the same story, augmentation technique, and prompting style. This multi-call approach enabled us to assess the stability and accuracy of each model's outputs under identical conditions, providing a solid basis for comparative analysis across different model setups.

5.1 Model Comparison

The *GPT-4-turbo* and *GPT-4o* model demonstrated similar high accuracies of 83.5% and 82.7%, respectively, showcasing their robust capabilities in handling reasoning tasks. The *GPT-4o-mini*, while smaller, achieved an accuracy of 74.1%, indicating its proficiency despite having fewer parameters. Figure 2 summarizes the accuracy of each model across different configurations, highlighting the comparable performance of *GPT-4-turbo* and *GPT-4o* due to their advanced reasoning and inference abilities.

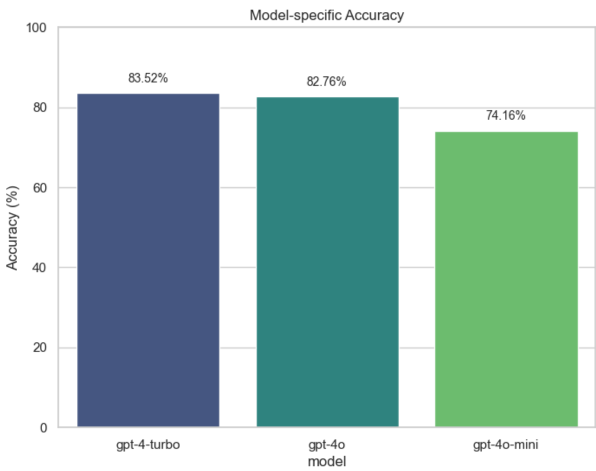

Figure 2: Accuracy comparison across models

5.2 Impact of Document Length on Model Accuracy

Figure 3 demonstrates how model accuracy is influenced by the number of pages in a document. The results indicate that *gpt-4o* and *gpt-4-turbo* exhibit strong resilience to increasing document lengths, maintaining consistent accuracy with only a minor decline as the number of pages grows. This suggests that these models are better equipped to handle long-context scenarios without significant performance degradation.

On the other hand, *gpt-4o-mini* shows a pronounced decline in accuracy as the number of pages increases. This steep drop-off highlights its limitations in processing and retaining information in longer documents. The disparity between *gpt-4o-mini* and the other models becomes more evident as the document length increases.

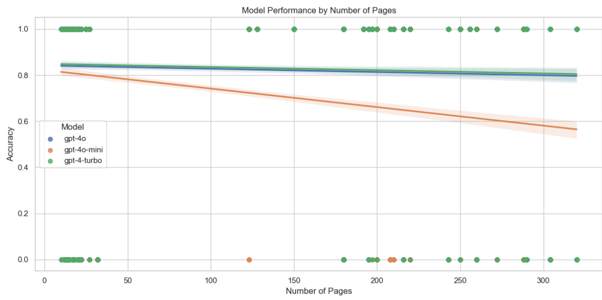

Figure 3: Accuracy distribution across the number of pages for different models.

5.3 Data Augmentation Analysis

The models achieved similar highest accuracy on the original text. However, when all character names were swapped, there was a noticeable drop in accuracy, suggesting that extensive alterations to familiar name patterns hinder the model's understanding of the narrative.

Interestingly, the accuracy increased for the *Harry Potter, Hollywood,* and *Bollywood* versions of the text, with the model performing similarly across these three cases. This indicates that the model benefits from contexts associated with well-known entities, possibly due to pre-training on a large corpus containing such references. Figure 4 summarizes the accuracy of each text variation, highlighting how character name familiarity and context influence model performance.

The table below specifies the meaning of different augmentation styles used in the analysis:

5.4 Prompting Technique Analysis

Prompting techniques had a notable impact on the model's ability to deduce the culprit's identity, with each method contributing differently to accuracy.

- **Normal Prompting**: As a baseline, normal prompting resulted in a relatively lower preci-

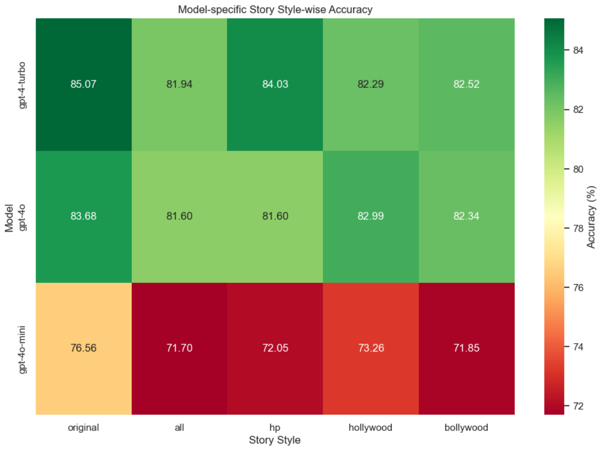

Figure 4: Accuracy across different data augmentation techniques.

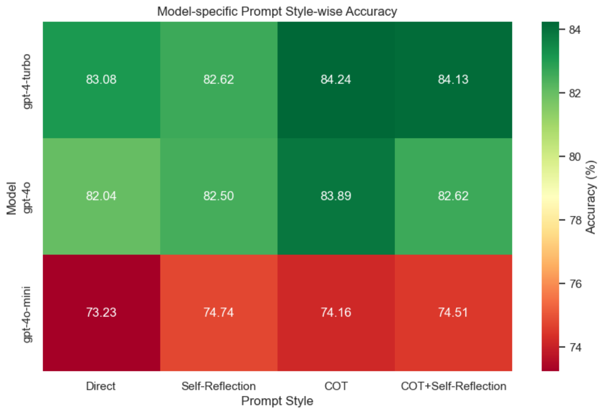

Figure 5: Accuracy across different prompting techniques.

Story Style	Description
original	Original text without any alterations.
all	All character names in the story swapped.
hp	Story with Harry Potter theme augmentation.
hollywood	Story augmented with a Hollywood theme.
bollywood	Story augmented with a Bollywood theme.

Table 1: Descriptions of different augmentation styles.

sion, as the model produced direct responses without deeper reasoning (Brown, 2020).

- **Self-Reflection Prompting**: Accuracy improved with Self-Reflection prompting, where the model refined responses through internal checks, leading to greater consistency in deductions (Shinn et al., 2024).

- **Chain-of-Thought (CoT) Prompting**: CoT prompting further increased accuracy by guiding the model through a structured reasoning process, allowing it to systematically address key narrative elements (Wei et al., 2022).

- **Chain-of-Thought + Self-Reflection (CoT + Self-Reflection)**: The combination of CoT and Self-Reflection yielded similar results as CoT, as the model generated logical step-by-step responses and then refined them, demonstrating the enhanced performance in narrative deduction (Madaan et al., 2024).

Figure 5 presents the accuracy achieved by each prompting technique, with substantial gains observed by adding CoT and Self-Reflexion, underscoring the effectiveness of combining structured reasoning and reflective validation.

Our results reveal that model architecture, data augmentation, and prompting techniques all play a significant role in shaping deductive performance. The findings highlight the crucial impact of structured prompting on enhancing model accuracy, particularly in complex narrative deduction tasks. These insights underscore the need for refined prompting strategies and comprehensive data preparation to optimize LLMs capabilities in inference-driven applications.

6 Conclusion and Future Work

We conclude by releasing our deductive reasoning capability benchmark, called WHODUNIT. We use this dataset to examine the deductive reasoning capabilities of large language models (LLMs) in complex narrative contexts, specifically focusing on mystery narratives that require nuanced inference and multi-step reasoning. Using a structured evaluation framework, we assessed the effects of model architecture, data augmentation, and various prompting techniques on the deductive accuracy of these LLM configurations — *GPT-4o, GPT-4-turbo,* and *GPT-4o-mini*. Our findings indicate that a combination of structured reasoning and reflective validation techniques, namely Chain-of-Thought and Self-Reflection prompting, significantly enhances model performance. Our results indicate that before a detective level reasonable understanding the models still have some progress to go in long-form narrative comprehension, and have to build robustness to changes in character names, while keeping the story plot intact. A key aspect of future work would be building long form comprehensive puzzle dataset, that would be able to test the limits of the LLM reasoning capabilities, and to reduce the impact of bias inducted during pre-training.

Limitations

This study is limited to short and medium-length stories due to the model's context length constraints, which restrict the analysis of longer narratives.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O Riedl. 2021. Automated storytelling via causal, commonsense plot ordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5859–5867.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jensen Huang and Ilya Sutskever. 2023. Interview with jensen huang and ilya sutskever. https://www.youtube.com/watch?v=GI4Tpi48DlA.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2022. Introducing chatgpt. https://openai.com/index/chatgpt/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

A Appendix

A.1 Extract from *A Case of Identity* by Arthur Conan Doyle

***Culprit:* James Windibank**

Point of Reveal:

"My dear fellow," said Sherlock Holmes as we sat on either side of the fire in his lodgings at Baker Street, "life is infinitely stranger than anything which the mind of man could invent. We would not dare to conceive the things which are really mere commonplaces of existence.

...

"Certainly," said Holmes, stepping over and turning the key in the door. "I let you know, then, that I have caught him!"

"What! where?" shouted Mr. Windibank, turning white to his lips and glancing about him like a rat in a trap.

"Oh, it won't do—really it won't," said Holmes suavely. "There is no possible getting out of it, Mr. Windibank. It is quite too transparent, and it was a very bad compliment when you said that it was impossible for me to solve so simple a question. That's right! Sit down and let us talk it over."

Our visitor collapsed into a chair, with a ghastly face and a glitter of moisture on his brow. "It—it's not actionable," he stammered.

...

As I expected, his reply was typewritten and revealed the same trivial but characteristic defects. The same post brought me a letter from Westhouse & Marbank, of Fenchurch Street, to say that the description tallied in every respect with that of their employe, James Windibank. Voila tout" "And Miss Sutherland?" "If I tell her she will not believe me. You may remember the old Persian saying, 'There is danger for him who taketh the tiger cub, and danger also for whoso snatches a delusion from a woman.' There is as much sense in Hafiz as in Horace, and as much knowledge of the world."

A.2 Extract from *Silver Blaze* by Arthur Conan Doyle

***Culprit:* John Straker**

Point of Reveal:

I am afraid, Watson, that I shall have to go," said Holmes, as we sat down together to our breakfast one morning. "Go! Where to?" "To Dartmoor; to King's Pyland."

...

"The real murderer is standing immediately behind you." He stepped past and laid his hand upon the glossy neck of the thoroughbred.

"The horse!" cried both the Colonel and myself.

"Yes, the horse. And it may lessen his guilt if I say that it was done in self-defence, and that John Straker was a man who was entirely unworthy of your confidence. But there goes the bell, and as I stand to win a little on this next race, I shall defer a lengthy explanation until a more fitting time."

...

My eyes fell upon the sheep, and I asked a question which, rather to my surprise, showed that my surmise was correct. "When I returned to London I called upon the milliner, who had recognised Straker as an excellent customer of the name of Derbyshire, who had a very dashing wife, with a strong partiality for expensive dresses. I have no doubt that this woman had plunged him over head and ears in debt, and so led him into this miserable plot." "You have explained all but one thing," cried the Colonel "Where was the horse?" "Ah, it bolted, and was cared for by one of your neighbours. We must have an amnesty in that direction, I think. This is Clapham Junction, if I am not mistaken, and we shall be in Victoria in less than ten minutes. If you care to smoke a cigar in our rooms, Colonel, I shall be happy to give you any other details which might interest you.

A.3 List of Stories and Authors

Type	Title	Author Name
Novel	A Study in Scarlet	Arthur Conan Doyle
Novel	Crime and Punishment	Fyodor Dostoevsky
Novel	Clouds of Witness	Dorothy L. Sayers
Novel	File No. 113	Emile Gaboriau
Novel	Find the Woman	G. K. Chesterton
Novel	Silver Blaze	Arthur Conan Doyle
Novel	That Affair Next Door	Anna Katherine Green
Novel	The Borough Treasurer	J. S. Fletcher
Novel	The Clue of the Twisted Candle	Edgar Wallace
Novel	The Crooked Man	Arthur Conan Doyle
Novel	The Crystal Stopper	Maurice Leblanc
Novel	The Curved Blades	Carolyn Wells
Novel	The D'Arblay Mystery	R. Austin Freeman
Novel	The Fellowship of the Frog	Edgar Wallace
Novel	The Hound of the Baskervilles	Arthur Conan Doyle
Novel	The Insidious Dr. Fu Manchu	Sax Rohmer
Novel	The Leavenworth Case	Anna Katherine Green
Novel	The Lerouge Case	Emile Gaboriau
Novel	The Man in Lower Ten	Mary Roberts Rinehart
Novel	The Man in the Brown Suit	Agatha Christie
Novel	The Murder of Roger Ackroyd	Agatha Christie
Novel	The Murder on the Links	Agatha Christie
Novel	The Mysterious Affair at Styles	Agatha Christie
Novel	The Mystery of the Blue Train	Agatha Christie
Novel	The Mystery of the Yellow Room	Gaston Leroux
Novel	The Opal Serpent	Fergus Hume
Novel	The Problem of Thor Bridge	Arthur Conan Doyle
Novel	The Secret Adversary	Agatha Christie
Novel	The Sign of the Four	Arthur Conan Doyle
Novel	The Teeth of the Tiger	Maurice Leblanc
Novel	The Unpleasantness at the Bellona Club	Dorothy L. Sayers
Novel	The Valley of Fear	Arthur Conan Doyle
Novel	Trent's Last Case	E. C. Bentley
Novel	Unnatural Death	Dorothy L. Sayers
Novel	Whose Body? A Lord Peter Wimsey Novel	Dorothy L. Sayers
Novel	X Y Z: A Detective Story	Anna Katherine Green
Short Story	A Case of Identity	Arthur Conan Doyle
Short Story	Silver Blaze	Arthur Conan Doyle
Short Story	The Adventure of Black Peter	Arthur Conan Doyle
Short Story	The Adventure of Charles Augustus Milverton	Arthur Conan Doyle
Short Story	The Adventure of Shoscombe Old Place	Arthur Conan Doyle
Short Story	The Adventure of the Abbey Grange	Arthur Conan Doyle
Short Story	The Adventure of the Beryl Coronet	Arthur Conan Doyle
Short Story	The Adventure of the Blue Carbuncle	Arthur Conan Doyle
Short Story	The Adventure of the Bruce-Partington Plans	Arthur Conan Doyle
Short Story	The Adventure of the Cardboard Box	Arthur Conan Doyle
Short Story	The Adventure of the Copper Beeches	Arthur Conan Doyle
Short Story	The Adventure of the Creeping Man	Arthur Conan Doyle

Short Story	The Adventure of the Dancing Men	Arthur Conan Doyle
Short Story	The Adventure of the Devil’s Foot	Arthur Conan Doyle
Short Story	The Adventure of the Dying Detective	Arthur Conan Doyle
Short Story	The Adventure of the Egyptian Tomb	Agatha Christie
Short Story	The Adventure of the Engineer’s Thumb	Arthur Conan Doyle
Short Story	The Adventure of the Empty House	Arthur Conan Doyle
Short Story	The Adventure of the Final Problem	Arthur Conan Doyle
Short Story	The Adventure of the Golden Pince-Nez	Arthur Conan Doyle
Short Story	The Adventure of the Illustrious Client	Arthur Conan Doyle
Short Story	The Adventure of the Mazarin Stone	Arthur Conan Doyle
Short Story	The Adventure of the Norwood Builder	Arthur Conan Doyle
Short Story	The Adventure of the Priory School	Arthur Conan Doyle
Short Story	The Adventure of the Red Circle	Arthur Conan Doyle
Short Story	The Adventure of the Second Stain	Arthur Conan Doyle
Short Story	The Adventure of the Six Napoleons	Arthur Conan Doyle
Short Story	The Adventure of the Solitary Cyclist	Arthur Conan Doyle
Short Story	The Adventure of the Speckled Band	Arthur Conan Doyle
Short Story	The Adventure of the Sussex Vampire	Arthur Conan Doyle
Short Story	The Adventure of the Three Gables	Arthur Conan Doyle
Short Story	The Adventure of the Three Garridebs	Arthur Conan Doyle
Short Story	The Adventure of Wisteria Lodge	Arthur Conan Doyle
Short Story	The Boscombe Valley Mystery	Arthur Conan Doyle
Short Story	The Disappearance of Lady Frances Carfax	Arthur Conan Doyle
Short Story	The Five Orange Pips	Arthur Conan Doyle
Short Story	The Hunter’s Lodge Case	Agatha Christie
Short Story	The Musgrave Ritual	Arthur Conan Doyle
Short Story	The Naval Treaty	Arthur Conan Doyle
Short Story	The Red-Headed League	Arthur Conan Doyle
Short Story	The Riddle of the Purple Emperor	Fergus Hume
Short Story	The Sturgis Wager: A Detective Story	Anna Katherine Green