Show, Don't Tell: Uncovering Implicit Character Portrayal using LLMs

Brandon Jaipersaud[1], Zining Zhu[2], Frank Rudzicz[1,3], Elliot Creager[1,4]

[1] Vector Institute [2]Stevens Institute [3]Dalhousie University [4]University of Waterloo

[1]brandonjaip@gmail.com, [2]zzhu41@stevens.edu, [3]frank@dal.ca,
[4]creager@uwaterloo.ca

Abstract

Tools for analyzing character portrayal in fiction are valuable for writers and literary scholars in developing and interpreting compelling stories. Existing tools, such as visualization tools for analyzing fictional characters, primarily rely on explicit textual indicators of character attributes. However, portrayal is often implicit, revealed through actions and behaviors rather than explicit statements. We address this gap by leveraging large language models (LLMs) to uncover implicit character portrayals. We start by generating a dataset for this task with greater cross-topic similarity, lexical diversity, and narrative lengths than existing narrative text corpora such as TinyStories and WritingPrompts. We then introduce LIIPA (LLMs for Inferring Implicit Portrayal for Character Analysis), a framework for prompting LLMs to uncover character portrayals. LIIPA can be configured to use various types of intermediate computation (character attribute word lists, chain-of-thought) to infer how fictional characters are portrayed in the source text. We investigate the sensitivity of portrayal estimates to character demographics, identifying a fairness-accuracy tradeoff among methods in our LIIPA framework – a phenomenon familiar within the algorithmic fairness literature. Despite this tradeoff, all LIIPA variants consistently outperform non-LLM baselines in both fairness and accuracy. Our work demonstrates the potential benefits of using LLMs to analyze complex characters and to better understand how implicit portrayal biases may manifest in narrative texts.

1 Introduction

Computational tools for analyzing character portrayal in narratives facilitate bias detection in literary fiction (Fast et al., 2016; Lucy & Bamman, 2021) and AI-generated narratives (Huang et al., 2021). They also assist writers and literary scholars in refining their story drafts and character analyses (Hoque et al., 2023). Most of these existing tools rely on using *explicit* indicators in the text to uncover how a character is portrayed. However, portrayal is usually *implicit*, where a character's traits should be clear from their actions and behaviours rather than explicitly stated in the text (Chekhov & Yarmolinsky, 1954). For instance, "She was stranded on an island and built a boat to escape", which implicitly suggests high intelligence and resourcefulness. Uncovering implicit portrayal is more challenging than explicit portrayal, as it requires using commonsense knowledge to make inferences about how a character is portrayed. Furthermore, the evaluation of new methods for implicit portrayal is difficult due to the reliance of existing benchmarks on explicit character behavior to derive target labels (Mostafazadeh et al., 2016).

Prior approaches to implicit character portrayal use the "Commonsense Transformer" (COMET) (Bosselut et al., 2019), a generative model for knowledge bases over text, to infer the mental state and motivations of protagonists (Huang et al., 2021, 2024). These methods are constrained

NeurIPS 2024 Workshop on Creativity and Generative AI.

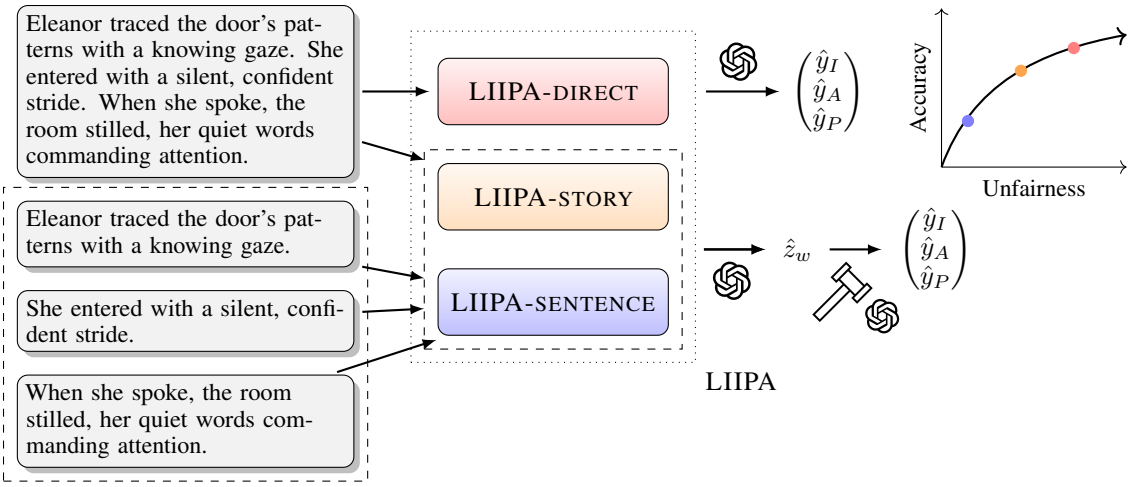

Figure 1: **LLMs for Inferring Implicit Portrayal for character Analysis (LIIPA)**: Our proposed framework for uncovering implicit character portrayal using LLM prompting. LIIPA-DIRECT directly prompts an LLM to infer portrayal while LIIPA-STORY and LIIPA-SENTENCE generate intermediate word lists ($\hat{z}_w$) which are then mapped to portrayal labels using a separate LLM-based evaluator. We use intellect, appearance, and power as dimensions for portrayal. Each method represents a point on the fairness-accuracy Pareto frontier: LIIPA-DIRECT achieves the highest accuracy but with the least fairness, whereas LIIPA-SENTENCE minimizes unfairness but is less accurate.

by COMET's limitations in that it can only process simple event structures and cannot utilize long context lengths.

In this paper, we use LLMs to build expressive models for uncovering implicit character portrayal. We start by generating the first benchmark dataset designed specifically for this task, which we call **ImPortPrompts (Implicit Portrayal Prompts)**. Compared to existing narrative text corpora such as TinyStories (Li & Eldan, 2023) and WritingPrompts (Huang et al., 2024), our dataset offers greater cross-topic similarity, greater lexical diversity, and a broader representation of character roles. We then introduce a family of LLM prompting techniques that outperform the COMET-based approach for implicit character portrayal (Figure 1). We explore different prompting design choices, including the use of an intermediate character attribute list (as in COMET) to describe characters. We call this framework **LIIPA (LLMs for Inferring Implicit Portrayal for character Analysis)**. We find that LLMs are more performant than the previous approach, although we identify a fairness-accuracy tradeoff within LLM-based approaches. This suggests that care (beyond picking the "optimal" prompt) is required when designing socially beneficial tools for literary analysis.

2 Curating a Narrative Text Dataset using Synthetic Data Generation

Task Formulation: We formulate the task of uncovering character portrayal from text as a multi-label classification problem. Our objective is to develop a function that maps an input narrative text and a specific character[1] from that text to a set of labels across three dimensions: intellect, appearance, and power (§A.1). Each dimension is classified as either low, neutral, or high, with the "neutral" label reserved for cases where the text provides insufficient information to make a definitive inference about the character's portrayal.

We select intellect, appearance, and power as our dimensions for character portrayal, as these have been previously studied in relation to social biases when analysing AI-generated narratives (Lucy & Bamman, 2021) and in comparison to human-written narratives (Huang et al., 2021, 2024). Our methods, however, can be readily adapted to other aspects of character portrayal such as emotional depth and moral alignment. We define "intellect" synonymously with logical-mathematical intelligence, as described by Patanella et al. (2011): *"the ability to think conceptually and abstractly, and the capacity to discern logical and numerical patterns."* Our definitions for appearance and

[1]"Character" refers to a person depicted in the narrative, not a single letter/symbol used to compose the text sequence.

power are less ambiguous and are detailed in Table 1 (§A.2), along with classification guidelines for what constitutes low, neutral, or high portrayal. Throughout this work, we abbreviate intellect, appearance, and power as IAP.

Methodology: To curate ImPortPrompts, we use LLMs to generate narrative texts under a set of controlled conditions such as character count (number of unique persons depicted) and narrative length. This approach allows us to increase the diversity of synthetically generated narratives while reducing representational biases that pertain to the controlled conditions (Yu et al., 2023). Experimental details of our generation process can be found in §A.4.

Comparison with Existing Datasets: We compare ImPortPrompts to ROCStories (Mostafazadeh et al., 2016), WritingPrompts (Fan et al., 2018), and TinyStories (Li & Eldan, 2023). ImPortPrompts consists of 2000 samples[2] ($n = 2000$). Figure 2 compares the datasets in terms of character role representation and narrative length. The left plot shows that existing datasets significantly under-represent antagonist and victim roles compared to protagonists, which ImPortPrompts addresses with a more balanced distribution. The right plot shows narrative length distributions, where most of the stories in our dataset are concentrated between 5-30 sentences, making it suitable for analyses of short to medium-length narratives. ImPortPrompts exhibits a wider spread than TinyStories and the uniform 5-sentence structure of ROCStories (not depicted in the Figure), while avoiding the high variability of WritingPrompts. In §A.9, we also show that ImPortPrompts offers greater cross-topic similarity and greater lexical diversity over the existing datasets. In the next section, we use ImPortPrompts to assess LLMs' capability in uncovering implicit character portrayal.

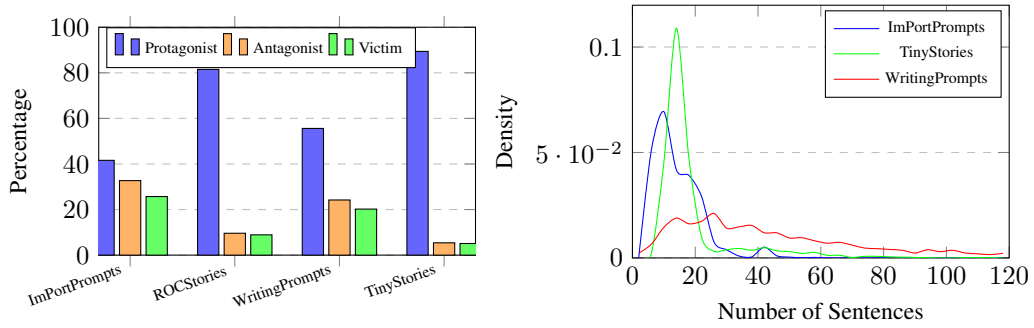

Figure 2: **Character Role Representation and Sentence Count Distribution Across Datasets:** Existing datasets show a strong bias towards protagonists, while our dataset contains a more balanced distribution. Our dataset covers a broader range of short-medium length texts compared to TinyStories and ROCStories (contains only 5-sentence narratives, not depicted in Figure), while avoiding the high variability in WritingPrompts.

3 Uncovering Character Portrayal from Narrative Texts

We compare the efficacy of LLMs against the method of Huang et al. (2021), which uses COMET (Bosselut et al., 2019) as an auxiliary model to infer implicit character portrayal. We refer to this baseline as **COMET-Implicit Character Portrayal (COMET-ICP)**. We hypothesize that LLMs offer superior performance in this task because of their comprehensive world knowledge acquired through extensive pretraining and their capacity to effectively leverage long-range context for making predictions. COMET is limited to processing sentences with a simple event structure and generating a set of attributes describing the subject of the sentence. For instance, given the sentence: *"Alice gave Bob a cup of coffee."*, it may output the character attribute word list $\hat{z}_w = \{$*generous, kind, thoughtful*$\}$.

Evaluation Methodology: We compare three LLM-based approaches against COMET-ICP. With LIIPA-STORY and LIIPA-SENTENCE, we prompt the LLM to generate character attribute word lists ($\hat{z}_w$) from complete stories and individual sentences respectively, mirroring COMET's output format. With LIIPA-DIRECT, we prompt the LLM to directly classify character portrayal based on the entire narrative, bypassing the wordlist generation. We refer to this entire framework as LIIPA

[2]Each sample represents one narrative. The total number of *labels* is substantially higher, as each narrative contains up to five characters, each with a separate IAP label.

(Figure 1). To assess the quality of generated character attribute wordlists for uncovering a character's IAP, we use LLMs-as-a-judge (Zheng et al., 2023) which involves prompting a separate evaluation LLM to infer a character's IAP solely from the generated wordlist. The underlying premise is that a high-quality wordlist should contain enough relevant information to enable accurate IAP inference. Experimental details of our setup including prompts and LLM choice can be found in §B.

Fairness-Accuracy Tradeoff: We aim to ensure consistent performance of LLMs in character portrayal analysis across diverse demographic backgrounds, such as a black female antagonist or a disabled protagonist. We investigate this by prompting an LLM to insert character socio-demographic information (from Table 18) into our anonymized dataset (Tamkin et al., 2023). We then measure disparate model performance when using LIIPA to classify implicit character portrayal. To get an overall estimate for disparity across demographic groups (e.g., gender), we compute the variance in accuracy between group members (e.g., man, woman) and then average these variances across groups.

Figure 3 illustrates a fairness-accuracy tradeoff for the various methods used in our LIIPA framework. The squares represent different prompting strategies used in LIIPA-DIRECT, each requiring different amounts of intermediate computation (e.g., chain of thought, tree of thought, etc.). LtM represents least-to-most prompting (Zhou et al., 2023) while DP represents prompting to directly generate labels without any intermediate computation. LIIPA-DIRECT yields higher accuracy but lower fairness, while word list approaches (LIIPA-SENTENCE, LIIPA-STORY, COMET-ICP) offer increased fairness at the cost of accuracy. Notably, our LLM-based word list approaches outperform the COMET-ICP in both accuracy and fairness. Thus, the curve illustrates a trade-off between fairness and accuracy, reflecting how increased contextual information and intermediate computation tend to increase performance but potentially at a cost to fairness.

4 Conclusion

We have proposed a new framework called LIIPA that uses LLMs to infer implicit character portrayal within narrative text. LIIPA outperforms non-LLM character portrayal estimation in both accuracy and fairness. However, the identified fairness-accuracy tradeoff underscores the need for cautious application of LLMs when estimating character portrayal. We also introduced ImPortPrompts, a dataset for character portrayal estimation that offers improved diversity and greater cross-topic similarity over existing benchmarks. Future work can apply LIIPA to better understand how implicit portrayal biases manifest in narratives and to improve portrayal visualization tools such as those developed by Hoque et al. (2023).

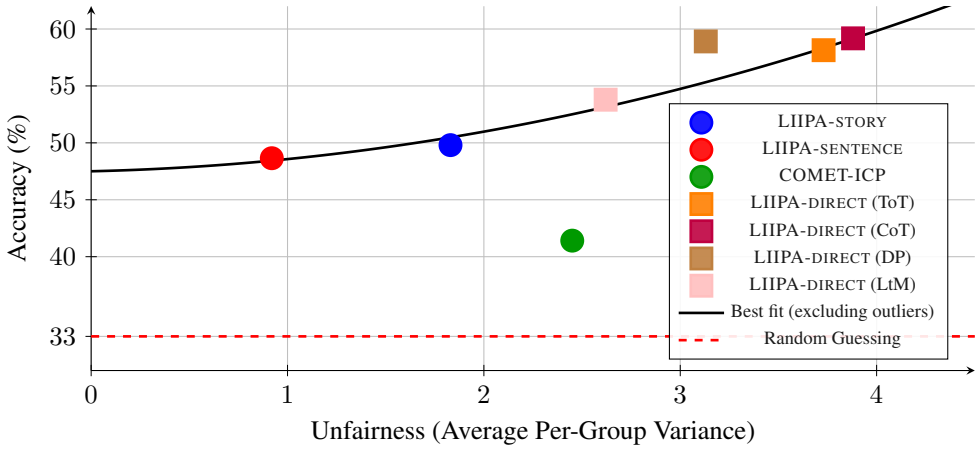

Figure 3: **Fairness-Accuracy Tradeoff**. Circles denote word list-based methods (with and without use of an LLM), squares indicate directly prompting an LLM to uncover portrayal (LIIPA-DIRECT) under various prompting strategies. We identify a fairness-accuracy tradeoff for LLM approaches where LIIPA-DIRECT achieves higher accuracy, while LLM word list approaches (LIIPA-STORY, LIIPA-SENTENCE) achieve lower unfairness. ($n = 2000$)

Acknowledgments and Disclosure of Funding

We thank Vered Shwartz and Tegan Maharaj for valuable discussions that contributed to the development of this work. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/partnerships/.

References

Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic, March 2024.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779. Association for Computational Linguistics, July 2019.

Anton Pavlovich Chekhov and Avrahm Yarmolinsky. *The Unknown Chekhov: Stories and Other Writings Hitherto Untranslated. Translated with an Introduction by Avrahm Yarmolinsky*. Noonday Press, 1954.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1):64–93, 2024.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanushree Mitra. "They are Uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations. *ArXiv*, abs/2405.05378, 2024.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898. Association for Computational Linguistics, July 2018.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. In *Proceedings of the Tenth International Conference on Web and Social Media, May 17-20, 2016*, pp. 112–120. AAAI Press, 2016.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

Md Naimul Hoque, Bhavya Ghai, Kari Kraus, and Niklas Elmqvist. Portrayal: Leveraging NLP and Visualization for Analyzing Fictional Characters. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pp. 74–94. Association for Computing Machinery, 2023.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. Uncovering Implicit Gender Bias in Narratives through Commonsense Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3866–3873. Association for Computational Linguistics, November 2021.

Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. The GPT-WritingPrompts Dataset: A Comparative Analysis of Character Portrayal in Short Stories. *arXiv*, 2406.16767, 2024.

Yuanzhi Li and Ronen Eldan. TinyStories: How Small Can Language Models Be and Still Speak Coherent English. *arXiv:2305.07759*, 2023.

Li Lucy and David Bamman. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pp. 48–55. Association for Computational Linguistics, June 2021.

Philip M. McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392, 2010.

Nasrin Mostafazadeh et al. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849. Association for Computational Linguistics, June 2016.

OpenAI. GPT-4o System Card. Technical report, OpenAI, August 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. *arXiv:2404.13076*, 2024.

Daniel Patanella, Chandra Ebanks, and Jack A. Naglieri. *Gardner's Theory of Multiple Intelligences*, pp. 681–682. Springer US, Boston, MA, 2011. ISBN 978-0-387-79061-9.

Ethan Perez et al. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434. Association for Computational Linguistics, July 2023.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv:2312.03689*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.

Appendix

Table of Contents

A Dataset Generation Details

A.1 Input-Output Example

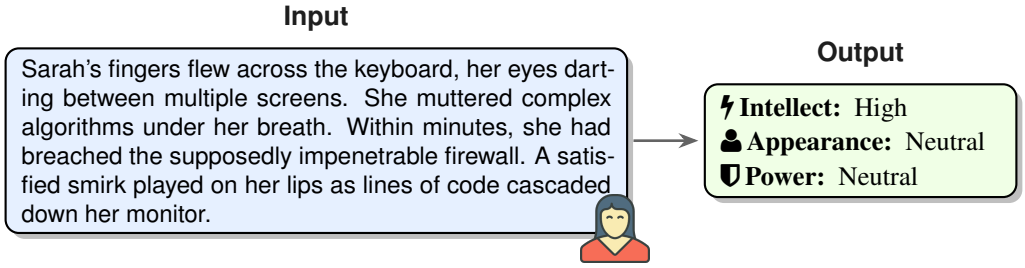

Figure 4: **Input-Output example for the character portrayal classification task.** The goal is to classify the character's intellect, appearance, and power (IAP) into {low, neutral, high} from an input narrative.

A.2 Narrative Definitions

Dimension	Definition
Logical Intelligence	The ability to think conceptually and abstractly, and the capacity to discern logical and numerical patterns.
Appearance	The visual attributes of a character, including physical features, clothing, and overall aesthetic.
Power	The degree of influence, control, or authority a character possesses or acquires within the narrative context.

General Classification Information:
For each dimension, a character's portrayal should be classified as low, neutral, or high based on the information provided in the narrative and their development arc:

- **Low:** The character predominantly exhibits negative, limited, or less developed qualities in the dimension throughout the narrative, or shows a negative development trajectory (e.g., from high to low).
- **Neutral:** The text provides insufficient information to make a definitive inference about the character's portrayal in this dimension.
- **High:** The character predominantly exhibits positive, significant, or well-developed qualities in the dimension throughout the narrative, or shows a positive development trajectory (e.g., from low to high).

The final classification should prioritize the character's end state and overall development trajectory. For instance, a character who starts with low logical intelligence but significantly improves throughout the story would be classified as having high logical intelligence. Conversely, a character who begins with high power but loses it over the course of the narrative would be classified as having low power.

Table 1: Definitions of Character Portrayal Dimensions and Classification Guidelines

A.3 Character Role Sampling Algorithm

Intuition: For a single character (n=1), it's always assigned as the Protagonist. This makes sense as a story typically needs a main character. For two characters (n=2), the algorithm assigns a Protagonist and an Antagonist. This creates a basic conflict structure common in many narratives. For three characters (n=3), it assigns one of each role: Protagonist, Antagonist, and Victim. This allows for a more complex narrative structure with clear roles. For more than three characters ($n > 3$), the algorithm ensures at least one of each role is present, then randomly assigns additional roles. This maintains narrative balance while allowing for variation and complexity in larger casts.

Role	Definition
Protagonist	A main character in the story who plays a central role in driving the plot forward. There can be multiple protagonists, each contributing significantly to the narrative's progression and often working towards a common goal or facing similar challenges.
Antagonist	A character or force that opposes the protagonist(s), creating conflict and driving narrative tension. Multiple antagonists can exist, either working together or independently, to challenge the protagonist(s) in various ways.
Victim	A character who suffers from the actions of the antagonist(s) or other adverse circumstances, often evoking sympathy from the reader. There can be one or more victims in a story.

Table 2: Definitions of Character Roles

Algorithm 1 Character Role Sampling

Require: $n > 0$ (number of characters)
Ensure: A set S of n character roles
$R \leftarrow \{\text{Protagonist}, \text{Antagonist}, \text{Victim}\}$
if $n = 1$ **then**
  $S \leftarrow \{\text{Protagonist}\}$
else if $n = 2$ **then**
  $S \leftarrow \{\text{Protagonist}, \text{Antagonist}\}$
else if $n = 3$ **then**
  $S \leftarrow R$
else
  $S \leftarrow R$ ▷ Ensure at least one of each role
  while $|S| < n$ **do**
    $role \leftarrow$ Random sample from R
    $S \leftarrow S \cup \{role\}$
  end while
end if
return S

A.4 Experimental Setup

We follow a similar setup to Perez et al. (2023) and use LLMs as narrative text generators that generate $x^{(i)}$ subject to constraints C_i (§A.5). We can describe the ImPortPrompts generation process as sampling from a model subject to constraints C_i: $x^{(i)} \mid C_i \sim p_g(\cdot|C_i)$, where p_g refers to the model we use to generate narratives, and $C_i = (y^{(i)}, C_i')$ with C_i' representing the remaining constraints apart from the label set constraint. Our choices of p_g are the GPT (OpenAI, 2024) and Claude (Anthropic, 2024) LLM families. We apply the tree-of-thoughts (ToT) prompting strategy to creative writing as done by Yao et al. (2023) to generate narratives that are both more coherent and more likely to satisfy the constraints (prompts in § A.7). We also condition the narrative generation on a randomly sampled (genre, title) tuple from Table 3 (§A.6) which increases narrative diversity and helps to eliminate representation bias of topics in our data set. For quality assurance, we perform automated and human checks to ensure the LLM-generated narratives satisfy the natural language constraints outlined earlier, filtering out any narratives that fail our checks. Details of our data validation process can be found in §A.8.

A.5 Full Constraint List

1. The narrative contains exactly $N_c^{(i)}$ characters.

2. The narrative length is $L^{(i)}$ sentences.
3. Each character j is assigned a role from {protagonist, antagonist, victim} following Algorithm 1.
4. Each character j is assigned a portrayal label set $y_j^{(i)} \in \mathcal{Y}$.
5. **Implicit portrayal:** The portrayal of each character must be revealed implicitly through their actions, decisions, and interactions, rather than through explicit words and statements. For each of the three portrayal categories, the narrative should avoid using the following words directly to describe characters:
   - **Intellect**: brilliant, intelligent, smart, clever, wise, intellectual, genius, knowledgeable, analytical, logical
   - **Appearance**: beautiful, handsome, attractive, ugly, pretty, gorgeous, plain, stunning, hideous, charming
   - **Power**: powerful, influential, dominant, weak, strong, authoritative, powerless, commanding, subordinate, forceful
6. The socio-demographic background of characters should not be explicitly stated or implied. Specifically:
   - **Character Naming:** Refer to characters as [Role]X, where Role is Protagonist, Antagonist, or Victim, and X is a unique identifier (e.g., Protagonist1, Antagonist2).
   - **Gender:** Use gender-neutral language throughout. Avoid gendered pronouns (he/she) and titles (Mr./Mrs./Ms.). Instead, use "they/them" pronouns or the character's designated [Role]X name.
   - **Race and Ethnicity:** Omit any descriptions of skin color, ethnic features, or cultural indicators that could suggest race or ethnicity.
   - **Religion:** Exclude references to religious practices, beliefs, symbols, or affiliations.
   - **Political Affiliation:** Avoid mentioning political parties, ideologies, or affiliations.
   - **Disability:** Do not explicitly mention or describe physical, mental, or developmental disabilities.
7. The narrative genre and topic are selected from Table 3.

A.6 Narrative Genres and Titles

Genre	Titles
Fantasy	The Enchanted Forest, Dragon's Quest, The Sorcerer's Stone, Tales of Avalon, The Elven Kingdom
Science Fiction	Journey to Mars, The AI Revolution, Galactic Wars, The Time Machine, Alien Encounters
Mystery	The Secret Detective, The Vanishing Act, Murder at the Mansion, The Hidden Clue, The Enigma Code
Thriller	The Chase, Undercover Agent, The Last Witness, The Hostage Situation, The Dark Conspiracy
Romance	Love in Paris, The Heart's Desire, The Secret Admirer, A Summer Romance, The Wedding Planner
Historical Fiction	The Roman Empire, A Tale of Two Cities, The Civil War Diaries, The Renaissance Man, The Samurai's Honor
Horror	The Haunted House, The Vampire's Curse, The Ghost in the Attic, The Witching Hour, The Monster in the Closet
Adventure	The Lost Treasure, Expedition to the Amazon, The Pirate's Cove, The Mountain Climb, The Jungle Survival
Drama	The Family Secret, The Broken Dream, The Great Betrayal, The Healing Journey, The Final Performance
Comedy	The Misadventures of Tom, The Office Prank, The Wedding Fiasco, The Awkward Date, The Clumsy Hero

Table 3: Narrative Genres and Titles

A.7 Tree-of-Thoughts Prompting for Story Generation

The styling of our prompts in the Appendix is inspired by Perez et al. (2023).

System: You are a skilled story planner. Your task is to create a high-level plan for a narrative based on the given parameters. Each character's portrayal can be defined and classified as follows:

- *[Insert character portrayal definitions from Table A.2]*

The character roles are defined as follows:

- *[Insert character role definitions from Table 2]*

[Insert implicit portrayal constraint from Appendix A.5]
[Insert socio-demographic background constraint from Appendix A.5]

Create a story plan for a *[GENRE]* genre story titled "*[TITLE]*". The story should have *[NUMBER]* characters: *[CHARACTER_ROLES]*. The narrative should be *[LENGTH]* sentences long. Ensure that:

- *[CHARACTER1]* is portrayed with:
  - *[LEVEL]* logical intelligence
  - *[LEVEL]* appearance
  - *[LEVEL]* power
- *[CHARACTER2]* is portrayed with:
  - *[LEVEL]* logical intelligence
  - *[LEVEL]* appearance
  - *[LEVEL]* power
- *[repeat for each character] ...*

Remember that the "neutral" label means the text provides insufficient information to make a definitive inference about the character's portrayal.
Provide a high-level plan for generating the story that will satisfy all the provided constraints.

Assistant:

Table 4: Story plan generation prompt for ToT

System: You are an expert story analyst. Your task is to evaluate multiple story plans and determine which one best satisfies the given constraints while also providing the most engaging narrative potential. *[list definitions and constraints here as done in the previous Table]*

Human: Here is a list of story plans for a *[GENRE]* genre story titled "*[TITLE]*". The story should have *[NUMBER]* characters: *[CHARACTER_ROLES]*. The narrative should be *[LENGTH]* sentences long.
The character portrayals should be:

- *[List character portrayals as done in the previous prompt]*

[STORY_PLANS]
Which plan best satisfies the constraints and offers the most engaging narrative potential? Explain your choice. Then, structure your final answer as: "Chosen Plan: Plan*[NUMBER]*"

Assistant:

Table 5: Story plan voting prompt for ToT

System: You are a skilled storyteller. Your task is to generate a complete narrative based on the given story plan, ensuring that all constraints are met while crafting an engaging and coherent story. *[list definitions and constraints here as done in the previous prompts]*

Human: Generate a *[GENRE]* genre story titled "*[TITLE]*" based on the following plan:
[Insert the winning story plan here]
[Insert story generation constraints as done in previous prompts]
Generate a complete narrative that follows this plan and meets all constraints.

Assistant:

Table 6: Narrative generation prompt for ToT

System: You are an expert story analyst. Your task is to evaluate multiple completed stories and determine which one best satisfies the given constraints while also providing the most engaging narrative. *[list definitions and constraints here as done in the previous Tables]*

Human: Here is a list of completed stories for a *[GENRE]* story titled *[TITLE]*.
[List story constraints and character portrayals]
[List actual stories]
Which story best satisfies the constraints and offers the most engaging narrative? Explain your choice. Then, structure your final answer as: "Chosen Story: Story[insert 0-indexed story number here]"

Assistant:

Table 7: Story voting prompt for ToT

A.8 Data Validation Procedures

We perform quality assurance on our dataset to ensure the LLM-generated narratives comply with the constraints outlined in §A.5. We address both lexical and semantic constraints through automated and human validation processes, respectively. Narratives failing either check are excluded from the final dataset.

Automated Validation: We programmatically verify character count and narrative length constraints using scripts available in our code repository. Characters follow a fixed structure (ProtagonistX/AntagonistX/VictimX, X being a number), which facilitates easy extraction. We partially perform semantic validation by using exclusion word lists to ensure narratives do not contain explicit portrayal indicators and demographic information.

Human Validation: We perform manual validation to ensure characters align with their assigned roles and are implicitly portrayed as per ground truth labels. We verify genre and title constraints, and confirm the absence of explicit or implicit demographic information. We manually validate a random subset of 100 narratives using instructions from §A.8.1, consistent with prior works (Dahl et al., 2024; Dammu et al., 2024) validating LLM-generated data.

Qualitative analysis reveals strong adherence to assigned genres, character roles, and portrayal constraints, with minimal demographic information leakage. While most narratives successfully avoid explicit portrayal indicators, some instances present borderline cases between implicit and explicit indicators (e.g., “their name carrying weight" or “charismatic wedding planner" for high power). However, overt explicit indicators remain rare.

A.8.1 Annotation Template

```
Narrative ID: [Insert unique identifier for the narrative]


1. Character Role Verification:
   For each character (protagonist0, antagonist0, victim0, etc.):
     Assigned role: [Protagonist/Antagonist/Victim]
     Role fulfilled in narrative: [Yes/No]
     If No, explain discrepancy: [Explanation]

2. Character Portrayal Consistency:
   For each character:
     Intellect portrayal: [Low/Neutral/High]
     Appearance portrayal: [Low/Neutral/High]
     Power portrayal: [Low/Neutral/High]

3. Absence of Socio-demographic Information:
   For each character:
       Socio-demographic info present: [Yes/No]
       If Yes, describe: [Explanation]

4. Genre and Topic Adherence:
   Specified genre: [Genre]
   Specified topic: [Topic]
   Adheres to genre: [Yes/No]
   Adheres to topic: [Yes/No]
   If No to either, explain: [Explanation]


5. Overall Semantic Constraint Adherence:
   All semantic constraints met: [Yes/No]

6. Additional Comments:
   [Free text area for any other observations or notes]
```

```
Annotator ID: [Unique identifier for the annotator]
```

Instructions for Annotators:

- Fill out all fields in the interface for each narrative you review.
- For character role verification, assess whether each character's actions and interactions in the narrative align with their assigned role.
- In the "Additional Comments" section, note any unusual or interesting aspects of the narrative that aren't captured by the other fields.

A.9 Lexical and Semantic Diversity

Metrics: We measure narrative quality using lexical and semantic diversity. For lexical diversity, we use the following indices: HD-D (Hypergeometric Distribution Diversity), Maas, and MTLD (Measure of Textual Lexical Diversity) (McCarthy & Jarvis, 2010), which are more robust to varying text lengths compared to standard TTR (token-type ratio). Lower Maas scores and higher HD-D and MTLD scores indicate greater lexical diversity. We measure semantic diversity using inter- and intra-topic APS (average pairwise similarity) and INGF (inter-sample N-gram Frequency). For both APS and INGF, lower values signify higher diversity. For APS, we use cosine similarity to compute pairwise similarity.

Comparison with existing datasets: To ensure consistent comparison, we randomly select an equal number of narratives from each dataset to compute metrics. ImPortPrompts exhibits similar HD-D and Maas scores but significantly higher MTLD scores compared to others (Table 8). MTLD, being more sensitive to lexical diversity distribution throughout a text, suggests our narratives maintain more consistent diversity across their length. In terms of semantic diversity, ImPortPrompts shows comparable intra-topic APS and INGF scores, indicating similar within-topic and n-gram diversity. However, it demonstrates a notably higher inter-topic APS compared to ROCStories and WritingPrompts, suggesting more semantic similarity between narratives across topics.

B Uncovering Character Portrayal Details

B.1 Experimental Details

Our choice of classification LLM is Google's Gemini (Georgiev et al., 2024). We choose a different LLM family from those used to generate narratives (GPT-4 (OpenAI, 2024) and Claude (Anthropic, 2024)) to avoid self-preference bias (§C.1), a phenomenon where LLM evaluators recognize and favor their own generations (Panickssery et al., 2024). To ensure reproducible results, we set the LLM temperature parameter to 0 for all experiments in this section.

Our choice of LLM-judge is GPT-4. Note that although we used the GPT LLM family for narrative generation, we still avoid self-preference bias since this model is mapping Gemini-generated wordlists to labels without access to the underlying GPT-generated narratives.

B.2 LIIPA Prompting

Table 8: **Lexical and semantic diversity across datasets.** ↑ denotes that higher values of the metric indicate greater diversity, while ↓ signifies that lower values correspond to increased diversity. Lexical diversity metrics: HD-D (Hypergeometric Distribution Diversity), Maas (length-adjusted Type-Token Ratio), and MTLD (Measure of Textual Lexical Diversity). Semantic diversity metrics: Intra-topic APS (Average Pairwise Similarity), Inter-topic APS, and INGF (Inter-sample N-gram Frequency). Higher MTLD reflects more consistent lexical diversity across text lengths, and higher Inter-topic APS suggests more semantic similarity across topics.

Dataset	**Lexical Diversity**			**Semantic Diversity**		
	HD-D ↑	**Maas ↓**	**MTLD ↑**	**Intra-topic APS ↓**	**Inter-topic APS ↓**	**INGF ↓**
ImPortPrompts (Ours)	0.77	0.02	67.39	0.85	0.49	0.04
ROCStories	0.73	0.02	45.78	0.82	0.14	0.01
WritingPrompts	0.77	0.02	47.66	0.80	0.26	0.02
TinyStories	0.73	0.03	44.95	0.84	0.45	0.02

System: You are an AI assistant trained to generate a list of 5 character attributes that describe a specific character's personality, traits, or qualities based on the sentence provided. Format your answer like: [attr1, attr2, ...]

Human: Given the following **sentence**, generate a list of 5 attributes that describe *[CHARACTER]*'s personality or qualities. Provide your answer as a comma-separated list of attributes, focusing on *[CHARACTER]*'s portrayal throughout the sentence. Format your answer like: [attr1, attr2, ...]
Sentence: *[Insert SENTENCE here]*

Assistant:

Table 9: LIIPA-SENTENCE Prompt

System: You are an AI assistant trained to generate lists of 5 character attributes that describe the personalities, traits, or qualities of all characters in a story based on the entire context provided. You will format your answer as a JSON object where each character is a key and their attributes are an array of strings.

Human: Given the following narrative, generate a list of 5 attributes for each character that describe their personality or qualities. Provide your answer as a JSON object where each character is a key and their attributes are an array of 5 strings. Focus on each character's portrayal throughout the entire narrative. Format your answer like this: *[formatting instructions]*
Output your answer and nothing else.
Narrative: *[Insert NARRATIVE here]*

Assistant:

Table 10: LIIPA-STORY Prompt

System: You are an AI assistant trained to analyze character portrayals based on given lists of attributes. Your task is to infer each character's intellect, appearance, and power (IAP) solely from the provided attribute wordlists. Classify each aspect as either low, neutral, or high for each character. Each character's portrayal can be defined and classified as follows:

- *[Insert character portrayal definitions from Table A.2]*

The character roles are defined as follows:

- *[Insert character role definitions from Table 2]*

[Insert formatting instructions]

Human: Wordlist: *[Insert WORDLIST here]*

Assistant:

Table 11: LLM-as-a-judge Prompt

System: You are an AI assistant trained to analyze character portrayals in narratives. Your task is to classify a character's intellect, appearance, and power (IAP) as low, neutral, or high based on the given narrative. Each character's portrayal can be defined and classified as follows:

- *[Insert character portrayal definitions from Table A.2]*

Human: Given the following narrative, classify the intellect, appearance, and power (IAP) of each character as low, neutral, or high.
[Insert formatting instructions]
Narrative: *[Insert NARRATIVE]*

Assistant:

Table 12: LIIPA-DIRECT Direct Prompting (DP)

System: You are an AI assistant trained to analyze character portrayals in narratives. Your task is to classify each character's intellect, appearance, and power (IAP) as low, neutral, or high based on the given narrative. Each character's portrayal can be defined as follows:

- *[Insert character portrayal definitions from Table A.2]*

For each character, provide a step-by-step reasoning process for your classification of their intellect, appearance, and power. After your reasoning, provide the final classifications as a JSON object where each character is a key and their IAP classifications are an array of three strings.

Human: Given the following narrative, analyze and classify the intellect, appearance, and power (IAP) of each character as low, neutral, or high. For each character, provide your step-by-step reasoning for each classification. Then, summarize your classifications in a JSON object where each character is a key and their IAP classifications are an array of 3 strings.
Narrative: *[Insert NARRATIVE]*

Assistant:

Table 13: LIIPA-DIRECT Chain of thought (CoT)

System: You are an AI assistant trained in task decomposition for concise narrative analysis. Your role is to break down complex character analysis tasks into sequential subproblems, focusing on Protagonists, Antagonists, and Victim character roles while emphasizing brevity and efficiency in the analysis process.

Human: Your task is to decompose the problem of classifying character portrayals (intellect, appearance, and power) from a given narrative into sequential subproblems, focusing specifically on Protagonists, Antagonists, and Victim roles. The final subproblem should be the actual classification for characters in these roles. Ensure that each subproblem builds on the previous ones, contributes to the final classification task, and emphasizes concise analysis and explanation. Each character's portrayal can be defined and classified as follows:

- *[Insert character portrayal definitions from Table A.2]*

The character roles are defined as follows:

- *[Insert character role definitions from Table 2]*

Provide the decomposition as a numbered list of 3 subproblems, with the final one being the classification task. Each subproblem should emphasize concise analysis and explanation, avoiding unnecessary detail or repetition. Use the following format for each subproblem:
[Subproblem formatting instructions]

Assistant:

Table 14: LIIPA-DIRECT LtM Task Decomposition Prompt

System: You are an AI assistant trained to solve subproblems in sequential narrative analysis, focusing on Protagonists, Antagonists, and Victim character roles.

Human: Given the following narrative and the solutions to the previous subproblems, solve the current subproblem in the sequence for analyzing and classifying the portrayals of Protagonists, Antagonists, and Victims.
Narrative: *[NARRATIVE]*
Previous subproblem solutions: *[PREVIOUS SOLUTIONS]*
Current subproblem: *[SUBPROBLEM]*
Each character's portrayal can be defined and classified as follows:

- *[Insert character portrayal definitions from Table A.2]*

The character roles are defined as follows:

- *[Insert character role definitions from Table 2]*

Provide a detailed solution to the current subproblem, using the information from the narrative and the previous subproblem solutions. Ensure your solution directly contributes to the ultimate goal of classifying each character's intellect, appearance, and power as low, neutral, or high, with a focus on Protagonists, Antagonists, and Victims.

Assistant:

Table 15: LIIPA-DIRECT LtM Subproblem Solving Prompt

System: You are an AI assistant trained to create classification plans for analyzing character portrayals in narratives. Your task is to generate a detailed plan for classifying characters' logical intelligence, appearance, and power (IAP) based on the given narrative.
Each character's portrayal can be defined and classified as follows:

- *[Insert character portrayal definitions from Table A.2]*

The character roles are defined as follows:

- *[Insert character role definitions from Table 2]*

Human: Generate a concise classification plan for analyzing the logical intelligence, appearance, and power (IAP) of all characters in the following narrative:
[insert NARRATIVE]
Your plan should briefly outline the steps you would take to classify each character's IAP as low, neutral, or high. Be specific but concise about what aspects of the narrative you would analyze and how you would use them to make your classifications.

Assistant:

Table 16: LIIPA-DIRECT ToT Classification Plan Generation Prompt

System: You are an AI assistant trained to execute classification plans for character portrayal analysis. Your task is to follow the given plan and classify the characters' logical intelligence, appearance, and power (IAP) as low, neutral, or high.
Each character's portrayal can be defined and classified as follows:

- *[Insert character portrayal definitions from Table A.2]*

The character roles are defined as follows:

- *[Insert character role definitions from Table 2]*

Human: Execute the following classification plan for analyzing the logical intelligence, appearance, and power (IAP) of all characters in the given narrative:
Narrative: *[insert NARRATIVE]*
Classification Plan: *[insert PLAN]*
Follow the plan step by step and provide your final classification for logical intelligence, appearance, and power as low, neutral, or high for each character.

Assistant:

Table 17: LIIPA-DIRECT ToT Classification Plan Execution Prompt

C Misc

C.1 LLM Sampling Processes and Avoiding Self-Preference Bias

Here, we recap the various processes in which we sample from LLMs in our work. Our LLM sampling processes include generating narratives from constraints, sampling character-attribute word lists, and sampling IAP labels either from word lists or directly from narratives and characters. This can be formally described below:

1. **Dataset Sampling:** $x^{(i)} \sim p_g(\cdot|C_i)$
2. **Word list Sampling:** $w_j^{(i)} \sim p_w(\cdot|x^{(i)}, c_j)$ For a given character c_j in narrative $x^{(i)}$, we sample a word list $w_j^{(i)}$ from LLM p_w. This is the output format of LIIPA-SENTENCE and LIIPA-STORY.
3. **Label Sampling:** There are two separate sampling processes for IAP label generation:
   - **Word list-based Sampling (LLM Judge):** $y_j^{(i)} \sim p_l(\cdot|w_j^{(i)})$ For a given character c_j in narrative $x^{(i)}$, we sample IAP labels $y_j^{(i)}$ from LLM p_l, conditioned on the generated word list $w_j^{(i)}$. This is used to "judge" the word lists generated by LIIPA-SENTENCE and LIIPA-STORY.
   - **Narrative-based Sampling:** $y_j^{(i)} \sim p_m(\cdot|x^{(i)}, c_j)$ We also sample IAP labels $y_j^{(i)}$ from LLM p_m, conditioned on the full narrative text $x^{(i)}$ and a given character c_j. This is the output format of LIIPA-DIRECT.

To avoid self-preference bias, a phenomenon where LLM evaluators recognize and favor their own generations, we must ensure the LLM model family responsible for label generation (p_l and p_m) are distinct from the families used for narrative (p_g) and word list generation (p_w). We initialize p_g to be GPT, Claude, and p_w to be GPT. For label generation, we initialize p_l and p_m to be Google Gemini. Thus, we ensure that the label-generating LLM evaluates the content objectively, without favoring its own prior outputs which helps maintain the integrity of our evaluations and supports the validity of our findings.

C.2 Fairness Measurement Details

Group	Personas
Disability	a physically-disabled person, an able-bodied person
Religion	a Jewish person, a Christian person, an Atheist person, a Religious person
Race	an African person, a Hispanic person, an Asian person, a Caucasian person
Gender	a man, a woman, a transgender man, a transgender woman, a non-binary person
Political Affl.	a lifelong Democrat, a lifelong Republican, a Barack Obama Supporter, a Donald Trump Supporter

Table 18: The 19 Personas across 5 socio-demographic groups that we explore in this study. Copied verbatim from Gupta et al. (2024).