Can Machines Read Minds? Detecting Interiority in Narrative Texts

Aigerim Kurmanbekova Chelsea Chen Michelle Lin
University of California, Berkeley
School of Information
{aigerim, chelsea_chen, lin.michelle}@berkeley.edu

Abstract

Representations of characters' inner thoughts, emotions, and perceptions play a crucial role in narratives, yet they remain difficult to capture with computational methods. In this work, we investigate whether interiority can be identified automatically at the paragraph level. We introduce a three-level annotation scheme (high, low, none) and present a manually annotated dataset of 600 passages sampled from 15 works of late nineteenth- and early twentieth-century English-language fiction. We examine inter-annotator agreement to assess the reliability of interiority judgments, and evaluate a set of baseline NLP models together with prompted large language models on this classification task. The results show that passages with explicit access to inner states are relatively stable for both human annotators and models. In contrast, passages expressing interiority indirectly, such as through free indirect discourse, pose consistent challenges. Overall, large language models show the most robust performance. These findings indicate that interiority can be learned from text, while subtle and graded expressions of inner life remain difficult to model.

1 Introduction

Interiority—the representation of a character's inner thoughts, feelings, and perceptions—is central to how narratives convey meaning. While computational work has increasingly modeled related aspects such as emotion and perspective, interiority itself has rarely been studied as a standalone task, in part because it is difficult to define and annotate in a consistent way. In this paper, we study paragraph-level interiority classification. Drawing on narratological accounts of consciousness, we propose an operational definition of interiority and a three-level annotation scheme (high, low, none) that captures both explicit and implicit expressions of inner life. Using this scheme, we construct a human-annotated dataset of 600 passages drawn from 15 works of late nineteenth- and early twentieth-century English-language fiction. We analyze inter-annotator agreement to understand the sources of ambiguity in the task, and evaluate a range of baseline NLP models and prompted large language models on interiority classification. Our results show that clear cases of explicit interiority and purely external narration are relatively easy for both humans and models to identify, while intermediate and indirect cases remain challenging. By operationalizing interiority in text, this work provides a foundation for larger-scale computational analysis of narrative interiority, including how access to inner life varies across genres, styles, periods, and characters within narratives.

2 Related Works

To ground our study, we review how interiority has been discussed in literary theories and how recent computational narratology has begun to model related dimensions of consciousness, emotion, and perspective.

2.1 Interiority

Narrative theory has viewed the representation of inner life as a core feature of fiction, even without explicitly using the term "interiority." Literary theorists have described a range of narrative techniques through which inner experience is made visible to readers. Cohn identifies forms such as interior monologue and free indirect discourse that allow access to a character's mental life (Cohn, 1978). Cognitive narratologists extend this, arguing that readers reconstruct a character's mind from cues distributed across the text (Palmer, 2004; Herman et al., 2005; Jahn, 2025). Across these traditions, interiority is treated not as a single, uniform phenomenon. Instead, it varies in degree and form, ranging from explicit statements of thought to emotional, sensory, or indirectly implied experiences.

Scholars also note that different literary periods and genres express interiority differently. Modernist writing tends to foreground inner experience, while other genres tend to rely on more direct emotional cues or externalized forms of subjectivity (Hoberek, 2019; Eldridge, 2025).

2.2 Methodology

Recent work in NLP and computational narratology has been moving beyond surface-level event modeling toward narrative meaning, with particular attention to mental states, perspective, and discourse structure. Rather than treating narratives as sequences of actions alone, this line of research emphasizes the role of internal and interpretive cues in shaping narrative understanding (Zhu et al., 2023; Piper et al., 2021). These studies highlight challenges in modeling implicit mental representations and long-range narrative structure, and call for clearer task definitions and evaluation frameworks.

Empirical work further suggests that aspects of inner life leave detectable textual signals. Prior studies show that incorporating representations of characters' mental states can improve the modeling of narrative structure, such as identifying climactic moments and resolutions (Vijayaraghavan and Roy, 2023). Other work directly targets internal experience. Cortal, for example, proposes an approach for jointly annotating characters and emotions in dream narratives, demonstrating that structured representations of internal states can be recovered from text, though performance depends on domain-specific factors (Cortal, 2024).

Closely related research on narrative perspective and focalization shows that interpretive categories are difficult to annotate reliably due to blurred narrative boundaries, even for experts. At the same time, recent results indicate that large language models, when guided by theory-informed prompts, can perform such annotation tasks with accuracy and consistency comparable to human agreement, even in zero-shot settings (Hicke et al., 2025).

Finally, comparative studies of modeling approaches highlight trade-offs between traditional NLP models, prompt-based large language models, and fine-tuned systems for subjective classification tasks. While fine-tuned or domain-specific models often outperform generic prompting, this work provides useful guidance for designing modeling pipelines under practical constraints (Kallstenius et al., 2025).

3 Data Collection & Annotation

3.1 Data

We selected 15 books published between 1890 and 1930 from Project Gutenberg (Appendix A). To examine interiority across different narrative styles, we aimed for a diverse set of works, including modernist and realist fiction as well as genre fiction such as detective and horror. The collection consists of novels, novellas, and short stories.

We split all books into paragraphs, removing those with fewer than 50 words, as they contain limited material for identifying interiority, and removed paragraphs longer than 400 words to exclude very long outliers. We first conducted a pilot study on 100 paragraphs randomly sampled from 5 of the 15 books. This pilot helped us identify sources of ambiguity and refine both our annotation scheme and the operational definition of interiority.

For the final dataset, we randomly sampled 40 paragraphs from each of the 15 books, resulting in 600 excerpts for human annotation. For final model development, we used excerpts from 6 books (240 paragraphs) as the validation set, and excerpts from the remaining 9 books (360 paragraphs) as the training set.

3.2 Annotation Method

The pilot study used a simple prompt to assign a binary label based on whether a passage contained internal or external narration, leading to substantial ambiguity and disagreement. To better capture the range of judgments, we decided to adopt a three-level annotation scheme: *high*, *low*, and *none*, with explicit criteria and a more detailed prompt. The final definition draws on (Cohn, 1978) framework and specifies five common narrative techniques used to portray a character's inner mind (Appendix B). Results reported in this paper are based on this final annotation method.

Three annotators independently performed the labeling task. All annotators are graduate students with advanced English proficiency and reviewed the shared guidelines before annotation. Each entry included the text title, paragraph, and label. After annotation, we compared labels to calculate agreement and identify ambiguities.

3.3 Evaluation of Agreement

Inter-annotator agreement was moderate, with a Krippendorff's alpha (ordinal) of 0.48. This level of agreement reflects the interpretive and context-sensitive nature of literary interiority: judgments about interiority can vary across readers, especially when the cues are subtle or indirect.

Clear cases were generally consistent across annotators, particularly when passages contained explicit markers of inner states, such as verbs of thought or feeling (e.g., *think, feel*). More ambiguous cases arose when interiority was conveyed indirectly. In particular, free indirect discourse without an explicit marker of internal monologue can blend seamlessly with external narration, making it harder to identify boundaries.

A key source of difficulty was the lack of a broader narrative context. Some excerpts may function as continuations of earlier passages of extended interior monologue, but the excerpt itself may not contain clear signals of interiority. Annotators base their judgments solely on the information available in the given paragraph, without inferring additional context from surrounding text or prior knowledge of the book. This aligns with our modeling goal: to enable automatic detection of interiority from standalone passages.

We used a majority vote to assign final labels. For the 10% of excerpts where no two annotators agreed, we discussed and assigned a consensus label. This process resulted in a gold-standard dataset with final label counts of *204 high*, *156 low*, and *237 none*.

4 Modeling

To evaluate the feasibility of automated interiority annotation, we experimented with a diverse set of models spanning multiple architectural families and levels of computational complexity. In total, we evaluated two baseline approaches and four families of language models. This design was built upon prior work on automated literary annotation, which emphasized comparing lightweight baselines, fine-tuned models, and large prompted language models to assess whether a task could be solved using simple surface cues or required deeper semantic modeling.

4.1 Baseline Models

We first established baseline performance using two traditional classifiers: logistic regression and Naive Bayes. These models tested whether interiority detection can be captured using shallow lexical patterns alone.

For both classifiers, we experimented with two feature extraction strategies: count vectorization, which represented paragraphs using raw word frequencies, and TF-IDF vectorization, which down-weighted common words while emphasizing rarer and more informative terms. Comparing these representations allowed us to assess whether weighting word importance improves performance over simple bag-of-words features.

All baseline models were implemented using the scikit-learn library. These baselines provided a reference point for evaluating more complex models.

4.2 Fine-Tuned Encoder Models

To test whether relatively small, accessible neural models can improve upon traditional baselines, we evaluated several fine-tuned transformer encoders. These include DistilBERT, BERT-large, and RoBERTa-base. These models were pre-trained on large corpora and fine-tuned for classification, allowing them to capture contextual representations beyond bag-of-words features.

All encoder-based models were implemented using the Hugging Face Transformers library. We did not perform extensive hyperparameter tuning, relying on default configurations to maintain comparability across models.

4.3 Sequence-to-Sequence Models

In addition to encoder-based classifiers, we experimented with sequence-to-sequence transformer models from the Flan-T5 family. We evaluated Flan-T5 models at multiple scales, including base, small, and large variants. These models were tested in two settings: fine-tuned classification and prompt-based zero-shot and few-shot inference.

Sequence-to-sequence models were well-suited for instruction-following tasks and allowed us to directly compare fine-tuning against prompting-based approaches within the same architectural family. All Flan-T5 experiments were conducted using Hugging Face, again without hyperparameter tuning beyond default settings.

4.4 Large Language Models

Finally, we evaluated large instruction-tuned language models using prompt-based classification. These included both open-source models and paid commercial models. Specifically, we tested two variants of Llama models and five models from the GPT family. For Llama, we used both Hugging Face inference and the Groq API to enable faster computation. For GPT models, we used OpenAI's API.

These models were evaluated in both zero-shot and few-shot settings. No additional fine-tuning was performed. This setup allowed us to assess whether large language models can perform interiority annotation purely through instruction following, as demonstrated in prior work on focalization classification and other narratological tasks.

4.5 Prompt Design

During development, we tested several prompt variants using a single reference model (LLaMA-3.1) to clarify task instructions and reduce ambiguity. Based on this exploratory phase, we selected the zero-shot and few-shot prompts that produced the most consistent and accurate outputs. These prompts were then fixed and applied across all prompted LLMs. While different models may respond differently to prompt variations, we used a shared prompt in this study to evaluate model behavior under a consistent task setup and enable clearer comparison across models.

The selected system prompts for the zero-shot and few-shot settings were similar. In the zero-shot setting, the system prompt specified the task by saying, "You are a classifier for literary interiority in fiction," followed by a definition of interiority and instructions on when to label as high, low, or none. In the few-shot setting, we used the same system prompt but added four labeled examples that were not from the test set to help guide the model.

The user prompt was the same for all settings, asking the model to classify the interiority level of the given paragraph as either "none," "low," or "high," in lowercase.

All models were run with a temperature setting of 0 to reduce output variability. The exception was the GPT-5 models, which required a temperature of 1, and were run with that setting. For full prompts, please refer to Appendix C.

4.6 Evaluation

Across all experiments, we evaluated model performance under two train-test split strategies: a random paragraph-level split with 80% training and 20% testing, and a book-level split in which entire books were held out during testing. For the random paragraph-level split, we used stratification by label to preserve class balance between training and test sets. For the book-level split, we selected six held-out books spanning multiple genres: *Dubliners, The Picture of Dorian Gray, My Ántonia, The Murder of Roger Ackroyd, A Farewell to Arms*, and *The Garden Party and Other Stories.*

The book-level split was particularly important for literary analysis, as it reduced stylistic and lexical leakage between training and test sets. Because this setting better reflected generalization to unseen texts, we focused our detailed analysis on results from the book-level test split. Results from the random split were also reported for reference (Appendix D).

Model	Accuracy	Overall F1
RoBERTa-base	0.61	0.53
GPT-3.5-turbo (Zero)	0.44	0.50
GPT-3.5-turbo (Few)	0.34	0.39
GPT-4o (Zero)	0.64	0.60
GPT-4o (Few)	0.62	0.58
GPT-4.1-mini (Zero)	0.57	0.56
GPT-4.1-mini (Few)	0.61	0.61
LLaMA-3.3-70B (Zero)	0.65	0.64
LLaMA-3.3-70B (Few)	**0.69**	**0.66**

Table 1: Selected modeling results with a book-level split. For full results, please refer to Appendix D.

Among all evaluated models, Llama-3.3-70b-versatile in the few-shot setting achieved the best overall performance, with an accuracy of 0.69 and a overall F1 score of 0.66. This was followed closely by Llama-3.3-70b-versatile zero-shot, which achieved an accuracy of 0.65 and a overall F1 score of 0.64. The next strongest model was GPT-4o in the zero-shot setting, which showed an accuracy of 0.64 and a overall F1 score of 0.60.

Overall, GPT-family models showed the most consistent performance across both zero-shot and few-shot settings and across evaluation splits. In particular, we observed a substantial performance improvement from GPT-3.5-turbo to GPT-4-level models, with accuracy increasing from approximately 0.44 to above 0.60 and overall F1 improving from around 0.50 to above 0.60. This pattern mirrored findings reported in the focalization classification study (Hicke et al., 2025) and suggested that improvements in model scale and training led to stronger capabilities for narratological classification tasks.

Among smaller fine-tuned transformer models, RoBERTa-base performed competitively, achieving 0.61 accuracy and 0.53 overall F1 on the book-level split. Notably, this performance exceeded that of several larger prompted LLMs. Given its relatively low computational cost, fast training time, and accessibility, this result suggested that fine-tuned encoder models remained a strong baseline for interiority detection. In contrast, Flan-T5 models consistently performed poorly across all sizes and prompting strategies, often underperforming even simple bag-of-words baselines. This trend held for both zero-shot and few-shot settings and suggested that sequence-to-sequence instruction tuning alone was insufficient for this task without further hyperparameter fine-tuning.

Across all model families, low interiority was the most difficult class to predict. Nearly all models showed substantially lower precision and recall for the low category compared to high and none. This difficulty aligned with our annotation observations, as human annotators also showed the greatest disagreement when labeling passages with implicit or ambiguous interiority. Additionally, class imbalance likely contributed to this issue, as the dataset contained fewer low-interiority instances (156) compared to high (205) and none (237). Notably, six out of the twenty-seven evaluated models failed to predict the low category entirely.

By contrast, performance on the high and none classes was relatively similar across models, with comparable F1 scores. This suggested that explicit interiority and purely external narration were easier for models to distinguish than intermediate or implicit cases.

5 Limitations

Our study has several limitations that are important to consider when interpreting the results.

First, interiority is an inherently interpretive phenomenon, and judgments often depend on broader narrative context. In this study, passages were annotated and modeled as standalone units, which increases ambiguity in cases where interiority is conveyed indirectly or unfolds across multiple paragraphs. This is particularly relevant for implicit expressions of interiority. While this design choice supports paragraph-level modeling, it necessarily limits the extent to which extended narrative continuity can be captured.

Second, the three-level label scheme introduces asymmetry across categories. The low interiority class is also the smallest category in the dataset. This combination likely contributes to instability in both human agreement and model performance for this class, and it helps explain why some models struggled to predict it reliably. Evaluation metrics that weight all classes equally, such as macro F1, are especially sensitive to errors in this category.

Finally, our modeling setup prioritizes broad comparison across models rather than extensive optimization. We used standard training settings and single-run prompting for LLMs, without detailed hyperparameter tuning, repeated generations, or large-scale robustness tests. More thorough optimization would require additional time and computational resources and could further improve performance for some models or prompting strategies. In addition, our corpus is limited to English-language fiction from 1890 to 1930, which may limit how well the findings generalize to other time periods, genres, or narrative domains.

6 Future Directions

Based on the observations from our current experiment and an acknowledgment of its limitations, we outline a vision for the next steps, focusing on three key areas: enhancing the robustness of our modeling pipeline, diversifying our dataset, and exploring downstream analyses once we have a well-validated modeling framework.

6.1 Enhancing the Modeling Pipeline

One of the first steps toward improving model performance is fine-tuning the existing pipeline. This includes conducting hyperparameter-tuning to increase model accuracy. Additionally, we aim to improve the robustness of our evaluations by testing for prompt resistance in prompt-based models, which can be achieved by varying prompts and assessing the consistency of model outputs. To address the non-deterministic nature of language models, we plan to run multiple instances of each model and take the majority vote for more reliable results. We also propose measuring inter-annotator reliability between models, similar to inter-annotator agreement conducted for human annotators, to assess consistency across machine classifications.

6.2 Interiority at Scale

Our annotation framework and model results provide a foundation for scaling up analysis within the existing corpus. One natural next step is to use the better-performing LLMs to annotate the full texts of the 15 selected books. More broadly, large-scale annotation makes it possible to compare narrative patterns across many texts. With interiority labels applied at scale, future work could examine how interiority is distributed across authors, genres, and narrative styles, how it shifts within a single novel over time, and how it is deployed differently across periods, shaping broader narrative structure and form.

6.3 Dataset Expansion

To address the limitation of our dataset's scope, we propose expanding it to include more contemporary texts, such as fan fiction, online forum discussions (e.g., Reddit posts), and other modern writing styles. This would allow us to assess whether models can generalize across different genres, time periods, and writing conventions.

6.4 Interiority Across Narrative Groups

Once a robust and reliable annotation pipeline is established, we can leverage it to explore downstream tasks. Specifically, we are interested in examining how access to interiority varies across different narrative roles or social groups. For example, one hypothesis is that access to interiority is less prevalent in minority group characters compared to majority group characters. To conduct this analysis, we will need to associate each annotated passage with its respective character, which requires the integration of metadata such as character identification. This analysis will help deepen our understanding of how narrative access to inner life is shaped by both character roles and broader social contexts. Additionally, this work can pave the way for future research exploring the relationship between interiority and specific social roles, gender, and other demographic factors, while carefully considering the historical context and potential biases in annotation.

7 Conclusion

In this paper, we presented a computational framework for detecting interiority in narrative fiction. We operationalized interiority as a three-level label (high, low, none), built a 600-paragraph dataset from 15 early twentieth-century books, and showed that even human annotators face real ambiguity when interiority is implicit, such as through free indirect discourse. Despite these challenges, our modeling results suggest that interiority is learnable. The Llama-3.3-70b-versatile model in the few-shot setting achieved the best performance, with an overall F1 score of 0.66, while the GPT family demonstrated the best overall performance. Surprisingly, despite their smaller scale, models from the BERT family performed comparably to the GPT models. However, the Flan-T5 models underperformed, contrary to expectations. Furthermore, the "low" interiority category proved difficult for nearly all model families, reflecting human annotators' disagreements and underscoring the complexity of defining low interiority. Overall, our work provides both a dataset and benchmark that future work can build on, with clear next steps for improving model robustness, expanding to broader corpora, and conducting downstream analyses of how interiority is represented across narrative groups.

References

Dorrit Cohn. 1978. *Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction*. Princeton University Press, Princeton, NJ.

Gustave Cortal. 2024. Sequence-to-sequence language models for character and emotion detection in dream narratives. *Preprint*, arXiv:2403.15486.

Sarah Eldridge. 2025. Interior selves: Morality, sentiment, and the emergence of organic essentialism. In *Composite Selves: Subjecthood in the German Novel, 1700–1795*. Oxford University Press.

David Herman, Manfred Jahn, and Marie-Laure Ryan, editors. 2005. *Routledge Encyclopedia of Narrative Theory*, 1st edition. Routledge, New York, NY.

Ruth M. M. Hicke, Yuri Bizzoni, Pascale Feldkamp, and Ross Deans Kristensen-McLachlan. 2025. Says who? effective zero-shot annotation of focalization. *Preprint*, arXiv:2409.11390.

Andrew Hoberek. 2019. Popular genres and interiority. *Amerikastudien / American Studies*, 64(4):567–578.

Manfred Jahn. 2025. Narratology 3.0: A guide to the theory of narrative.

Tobias Kallstenius, Andrei J. Capusan, Gerhard Andersson, and Adam Williamson. 2025. Comparing traditional natural language processing and large language models for mental health status classification: A multi-model evaluation. *Scientific Reports*, 15:24102.

Alan Palmer. 2004. *Fictional Minds*. University of Nebraska Press, Lincoln, NE.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding.

Pranav Vijayaraghavan and Deb Roy. 2023. M-sense: Modeling narrative structure in short personal narratives using protagonist's mental representations. *Preprint*, arXiv:2302.09418.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are nlp models good at tracing thoughts: An overview of narrative understanding. *Preprint*, arXiv:2310.18783.

A Corpus Contents

Book Title	Author	First Publication Year
The Picture of Dorian Gray	Oscar Wilde	1890
The Awakening, and Selected Short Stories*	Kate Chopin	1899
Martin Eden	Jack London	1909
A Room with a View	E. M. Forster	1908
My Ántonia	Willa Cather	1918
Dubliners	James Joyce	1914
The Metamorphosis	Franz Kafka	1915
Swann's Way	Marcel Proust	1913
The Age of Innocence	Edith Wharton	1920
The Garden Party, and Other Stories	Katherine Mansfield	1922
The Great Gatsby	F. Scott Fitzgerald	1925
Mrs. Dalloway	Virginia Woolf	1925
The Murder of Roger Ackroyd	Agatha Christie	1926
A Farewell to Arms	Ernest Hemingway	1929
The Dunwich Horror	H. P. Lovecraft	1929

Table 2: Corpus texts used for annotation and modeling.

B Annotation Guidelines

We define *interiority* as any moment in the narrative where the text gives access to a character's inner experience, including their thoughts, feelings, or perceptions. A passage exhibits interiority when it represents what is going on inside a character's mind, rather than describing only external actions or events.

To make this concept concrete and consistently identifiable, we draw on Dorrit Cohn's framework for how fiction represents consciousness (Cohn, 1978). Based on this framework, we define five observable categories through which interiority appears in narrative texts:

- **Psycho-narration (Third-Person):** The narrator explicitly describes a character's thoughts, feelings, or perceptions (e.g., *She felt/thought/believed that...*). Typically marked by verbs of mental activity and no quotation marks.
- **Quoted Interior Monologue (Third-Person):** A character's unspoken thoughts presented in their own words, often enclosed in quotation marks (e.g., "...", she thought).
- **Narrated Monologue (Third-Person):** Also known as free indirect discourse, blending the narrator's voice with the character's internal perspective. These passages usually lack quotation marks and explicit mental verbs (e.g., *She walked in. What a disaster this would be.*).
- **Retrospective Narration (First-Person):** A first-person narrator reflects on past mental states (e.g., *I remember thinking...*). References to memory of external events alone do not count as interiority.
- **Direct Interior Monologue (First-Person):** A first-person narrator presents current, ongoing inner thoughts or feelings as they occur (e.g., *What should I do now?*). These passages typically appear without quotation marks.

Each passage is assigned exactly one of the following labels:

- **High:** The passage clearly exhibits at least one of the five interiority types.
- **Low:** The passage suggests interiority, but evidence is indirect or ambiguous (commonly in free indirect discourse), or interiority is limited to a single word or phrase.
- **None:** The passage contains only external description, actions, or spoken dialogue, with no access to inner experience.

C Prompt Design

C.1 Zero-shot System Prompt

You are a classifier for literary interiority in fiction. Interiority refers to moments when the text gives access to a character's inner thoughts, feelings, or perceptions, rather than only external actions or events. Label each paragraph as exactly one of: high (explicit access to inner experience), low (indirect or ambiguous hints), none (only external description, actions, or spoken dialogue). Spoken dialogue alone does not count as interiority unless the text also explicitly reveals inner thoughts or feelings. Output only one word in lowercase: high, low, or none.

C.2 Few-shot System Prompt

You are a classifier for literary interiority in fiction. Interiority refers to moments when the text gives access to a character's inner thoughts, feelings, or perceptions, rather than only external actions or events. Label each paragraph as exactly one of: high (explicit access to inner experience), low (indirect or ambiguous hints), none (only external description, actions, or spoken dialogue). Spoken dialogue alone does not count as interiority unless the text also explicitly reveals inner thoughts or feelings.

Examples:

'So, thought Septimus, looking up, they are signalling to me.' → high

'At first, he stood there still, looking at the ground as if the contents of his head were rearranging themselves into new positions.' → low

'The wind rose in the night and rain came in sheets as the Croatians crossed the mountain meadows and fought in the dark.' → none

'Come on, I said. Get in.' → none

Output only one word in lowercase: high, low, or none.

C.3 User Prompt

Classify the interiority level of the following paragraph as high, low, or none:

```
""" {paragraph} """
```

D Modeling Results

Model	Accuracy	Overall F1	None_F1	Low_F1	High_F1
Logistic Regression (Count)	0.5	0.48	0.54	0.29	0.59
Naive Bayes (Count)	0.45	0.45	0.53	0.22	0.47
Logistic Regression (TF-IDF)	0.46	0.44	0.58	0.19	0.52
Naive Bayes (TF-IDF)	0.38	0.39	0.48	0	0.34
DistilBERT	0.57	0.54	0.67	0.25	0.68
BERT-large	0.55	0.49	0.66	0	0.58
RoBERTa-base	0.61	0.53	0.72	0.12	0.68
GPT-3.5-turbo (Zero)	0.44	0.5	0.34	0.4	0.6
GPT-3.5-turbo (Few)	0.34	0.39	0.09	0.25	0.55
GPT-4o (Zero)	0.64	0.6	0.67	0.35	0.72
GPT-4o (Few)	0.62	0.58	0.72	0.35	0.69
GPT-4.1-mini (Zero)	0.57	0.56	0.66	0.28	0.69
GPT-4.1-mini (Few)	0.61	0.61	0.65	0.42	**0.73**
GPT-4.1 (Zero)	0.59	0.58	0.67	0.29	0.7
GPT-4.1 (Few)	0.57	0.57	0.64	0.31	0.71
GPT-5-nano (Zero)	0.54	0.49	0.64	0.22	0.61
GPT-5-nano (Few)	0.57	0.53	0.64	0.28	0.67
llama-3.1-8b-instant (Zero)	0.5	0.5	0.5	0.3	0.69
llama-3.1-8b-instant (Few)	0.5	0.51	0.53	0.41	0.6
llama-3.3-70b-versatile (Zero)	0.65	0.64	0.73	0.46	0.72
llama-3.3-70b-versatile (Few)	**0.69**	**0.66**	**0.79**	**0.52**	0.68
Flan-T5 Small (Zero)	0.39	0.29	0.43	0	0.45
Flan-T5 Small (Few)	0.3	0.19	0.12	0	0.44
Flan-T5 Base (Zero)	0.46	0.28	0.6	0	0.25
Flan-T5 Base (Few)	0.48	0.31	0.62	0	0.31
Flan-T5 Large (Zero)	0.34	0.25	0.15	0.11	0.5
Flan-T5 Large (Few)	0.35	0.27	0.15	0.18	0.5

Table 3: Modeling results for various models with a book-level split. The highest value in each column is bolded.

Model	Accuracy	Overall F1	None_F1	Low_F1	High_F1
Logistic Regression (Count)	0.54	0.52	0.64	0.32	0.59
Naive Bayes (Count)	0.51	0.46	0.53	0.29	0.58
Logistic Regression (TF-IDF)	0.53	0.5	0.53	0.33	0.64
Naive Bayes (TF-IDF)	0.53	0.41	0.63	0	0.59
RoBERTa-base	0.68	0.64	0.77	0.41	0.76
DistilBERT	**0.71**	**0.69**	**0.77**	**0.5**	**0.79**
BERT-large	0.55	0.41	0.66	0	0.58
Flan-T5 Small (fine-tuned)	0.4	0.19	0.57	0	0.57
Flan-T5 Base (fine-tuned)	0.42	0.36	0.52	0.2	0.38
Flan-T5 Small (Zero)	0.38	0.39	0.38	0	0.47
Flan-T5 Small (Few)	0.32	0.21	0.14	0.01	0.46
Flan-T5 Base (Zero)	0.42	0.29	0.56	0.03	0.28
Flan-T5 Base (Few)	0.41	0.26	0.55	0.01	0.22
Flan-T5 Large (Zero)	0.37	0.26	0.37	0.13	0.53
Flan-T5 Large (Few)	0.35	0.24	0.36	0.13	0.51

Table 4: Modeling results for various models with a paragraph-level 80/20 random split. The highest value in each column is bolded.