Says Who? Effective Zero-Shot Annotation of Focalization

Rebecca M. M. Hicke[1,2] ⓘ, Yuri Bizzoni[2] ⓘ, Pascale Feldkamp[2] ⓘ, and
Ross Deans Kristensen-McLachlan[2,3,4] ⓘ

[1] Department of Computer Science, Cornell University, Ithaca, NY, USA
[2] Center for Humanities Computing, Aarhus University, Aarhus, Denmark
[3] Department of Linguistics, Cognitive Science, and Semiotics, Aarhus University, Aarhus, Denmark
[4] TEXT - Center for the Contemporary Cultures of Text, Aarhus University, Aarhus, Denmark

Abstract

Focalization describes the way in which access to narrative information is restricted or controlled based on the knowledge available to knowledge of the narrator. It is encoded via a wide range of lexico-grammatical features and is subject to reader interpretation. Even trained annotators frequently disagree on correct labels, suggesting this task is both qualitatively and computationally challenging. In this work, we test how well five contemporary large language model (LLM) families and two baselines perform when annotating short literary excerpts for focalization. Despite the challenging nature of the task, we find that LLMs show comparable performance to trained human annotators, with GPT-4o achieving an average F1 of 84.79%. Further, we demonstrate that the log probabilities output by GPT-family models frequently reflect the difficulty of annotating particular excerpts. Finally, we provide a case study analyzing sixteen Stephen King novels, demonstrating the usefulness of this approach for computational literary studies and the insights gleaned from examining focalization at scale.

Keywords: focalization, computational literary studies, large language models, immersivity

1 Introduction

Narratology and narrative theory provide rich frameworks of complex textual phenomena which can be used to explain how narrative discourse is ordered and how readers process text [17; 21; 43]. Since its inception, narrative theory has been applied to a wide variety of domains, from news-writing [36] to political discourse [46]. One particularly central notion is that of *focalization*, or the way in which information is constrained by the narrator's knowledge or perspective [19]. While focalization is often correlated with point-of-view, these phenomena are subtly different. Point of view (or perspective) describes the way a story is represented from a particular position or world view [35]. Focalization, on the other hand, refers more narrowly to the way in which a reader's access to narrative information is selected or restricted based on the knowledge of a narrator or character in the text [34]. In other words, focalization takes into account not only who (or what) is narrating, but also how that narrative voice is situated in the context of the story [45] – i.e., how information is disclosed. Analyses of focalization have shown promise in qualitative research for understanding both narrative structures and the impact of texts on readers.[1] Existing

[1] While we focus on literature here, focalization is also used to describe film and other media [16].

research also demonstrates how narrative focalization affects readers' character identification and empathy, as well as their experience of immersivity in a narrative [4; 10; 25]. As such, focalization both provides information on how narratives are constructed and serves as an intermediate step toward understanding more complex textual phenomena, such as affective reader response and identification. Moreover, the concept's relevance extends beyond literary narratives: patterns of focalization may also inform how affective framing operates in news discourse or political communication, where perspective and information control shape readers' emotional and ideological engagement [37; 42].

Recently, large language models (LLMs) have shown great promise for automating the annotation of many syntactic and semantic linguistic features [22; 47; 49]. These automated annotations allow researchers to study linguistic and literary phenomena on previously infeasible scales. In addition, examining the ability of LLMs to perform these challenging tasks provides researchers with a fuller understanding of their capabilities, particularly their strengths and weaknesses when performing nuanced annotations of complex, real-world texts.

This paper therefore studies LLMs' ability to annotate literary texts for focalization. We consider scenarios in which there is little human-annotated data available in an effort to make this methodology accessible to those without the time, money, or expertise to produce such annotations. Specifically, we evaluate the ability of LLMs from five model families — DistilBERT, RoBERTa, Flan-T5, Llama, and GPT — and two baselines — logistic regression and Naive Bayes — to annotate excerpts from sixteen novels by Stephen King for focalization mode. We find that GPT-4o with a zero-shot prompt achieves high (F1 = $\sim$85%) agreement with human consensus labels. Further, we determine that GPT-4o provides similar labelings when prompted multiple times on the same dataset and that it is resilient to prompt perturbations. Finally, we show that confidence values calculated from the log-probabilities output by GPT models frequently reflect the difficulty of annotating particular passages.

In addition, we demonstrate the usefulness of automated focalization annotations by analyzing the structure and flow of 16 novels by Stephen King, highlighting three outliers in this small corpus of his work. Considering the link between focalization and immersion in stories [10], we compare these annotations with measures of sensory information in each novel, finding that King appeals primarily to different senses across different focalization modes.

2 Related Works

Considerable recent work in NLP and computational humanities has examined the ability of language models to perform a variety of text-annotation tasks. Much of this research has fine-tuned models for tasks like co-reference annotation [22], classifying legal reasoning [49], distinguishing between historical and contemporary novels [9], recognizing spatial entities [26], and more [8]. With the advent of instruction-tuned LLMs like GPT and Llama, studies have further probed the ability of models to perform annotations in a zero or few-shot setting for tasks including identifying aspects of poetic form [51], story-telling [5], character roles [48], familial relationships [38], sentiment [40], genre [28], and more. While model performance varies across tasks, research has generally found that LLMs have a remarkable ability to perform nuanced literary and linguistic annotations even when given no or few examples. We build on this research by further exploring the capabilities of smaller, fine-tuned, and large, prompted LMs to annotate for focalization.

Despite the limitations of the term and its application [33], focalization has been useful in narrative analysis across both fiction and nonfiction texts. Generally, studies tend to adopt either the original taxonomy of Genette [19], or a revised approach such as that of Bal and Boheemen [7] or Chatman [12]. In the original taxonomy, Genette [19] distinguishes between three main types of focalization: internal, external, and zero. The latter corresponds to what is also generally known as an omniscient narrator – i.e., a type of narrator with knowledge of all events, as well as

Title	# Excerpts
The Gunslinger	430
The Girl Who Loved Tom Gordon	516
Dolores Claiborne	753
The Eyes Of The Dragon	791
Misery	821
Cujo	907
The Green Mile	967
The Dead Zone	1,006
Firestarter	1,103
Salem's Lot	1,163
The Waste Lands	1,298
Desperation	1,436
Insomnia	1,851
Wizard And Glass	1,885
Needful Things	1,898
The Stand	3,295

Table 1: Total number of automatically annotated excerpts per book (cf. Section 5).

of the characters and their thoughts. In internal focalization, narration is restricted to what a single character knows, hears, or sees, while in external focalization, it is based on the outside observation of characters or events. Significantly, modes of focalization are not equivalent to narrative point-of-view; third-person narration in particular is frequently internally focalized. Indeed, this is a core feature of so-called *free indirect style,* an important stylistic phenomenon in the history of English literature [41].

One of the major points of discussion since the term's introduction has been the distinction between external and zero focalization [34]. Further studies on focalization have demonstrated that it is a significant aspect of how readers process texts and how effects, such as suspense, affect, and identification, are generated in narrative text [4; 25].

Focalization was additionally mentioned as a task of interest in two recent surveys on computational narrative understanding [39; 44], although there have been few attempts to automate its annotation. Researchers in narrative generation have proposed frameworks that incorporate focalization [2; 6] and some work has attempted to identify similar concepts such as stream-of-consciousness [30] and types of speech (e.g. free indirect) [11]. However, none of these works use large, generative LMs or attempt to annotate directly for focalization. We are not aware of any existing research on automating focalization annotations at scale.

3 Methods

3.1 Data

We study a corpus of 16 novels by Stephen King included in the Chicago Corpus[2] and published between 1975 and 1999 (listed in Appendix A). We focus on novels by King because he is known to use multiple modes of focalization within a novel, often to create suspense [13]. Moreover, King is known as an extraordinarily productive and versatile author, publishing over a long period and across genres connected to differing registers and narrative strategies [15; 23; 27]. We may therefore expect King's use of focalization to vary across texts. In addition, studying the works of a single author allows us to determine whether focalization annotations can be used to identify structural outliers within a single author's oeuvre.

---

[2] `https://textual-optics-lab.uchicago.edu/us_novel_corpus`

```
Please state how the given novel excerpt is focalized, or the perspective
through which narrative information is provided. There are three modes of
focalization:

    • INTERNAL: From the perspective of a particular character. Describes
      what the character knows, including their thoughts, behaviors, and
      emotions.
    • EXTERNAL: From the perspective of an outside narrator. Describes
      characters' actions, behaviors, and settings. Communicates physically
      observable facts.
    • ZERO: From every perspective. The narration has an omniscient point of
      view and can describe the thoughts, behaviors, emotions, and actions of
      any character.

Speech in quotation marks counts as external focalization. Only respond with
one word representing the mode of focalization.

EXCERPT: ——

MODE:
```

Figure 1: The prompt used for all zero-shot experiments. The definitions for each mode of focalization are adapted from those given in Lijuan Chen and Lv [29]. The paragraph to be analyzed is inserted after the `EXCERPT:` heading.

To prepare data for training and evaluating models, we split each novel into paragraphs and retain only excerpts of at least 50 words. We choose this method of segmenting the novels because paragraphs are salient textual units with breaks selected by the author and are thus narratively meaningful. In addition, studying multiparagraph textual units may lead to sections with shifting focalization being falsely labeled as zero focalization. For human annotation and validation, we draw sixteen excerpts from each novel, creating a dataset of 256 paragraphs in total. We then sample 50 further excerpts from six novels[3] to create a minimal training dataset of 300 paragraphs. When annotating entire novels, we remove all excerpts containing front and back matter from the dataset. The remaining number of paragraphs annotated from each novel is displayed in Table 1.

3.2 Human Annotation

Each of the 256 excerpts in the evaluation dataset was annotated for focalization by three annotators. The first 96 samples were annotated by two native English speakers and one fluent speaker. Two of the annotators had post-secondary degrees in literature. The remaining 160 annotations were made by the two original annotators with post-secondary degrees in literature and by a third annotator, a fluent English speaker with a post-secondary degree in literature. In both cases, the annotators were instructed to work independently and did not have access to each other's annotations. Points of confusion were intentionally not discussed until after the initial annotation phase to produce a realistic measure of disagreement between informed human annotators.

For guidance, the annotators were given the same prompt as the LLMs (Figure 1). The prompt includes a basic definition of focalization and short descriptions of each focalization mode, adapted from those in Lijuan Chen and Lv [29]. One sentence in the prompt — "Speech in quotation marks counts as external focalization." — was only added after the first 96 excerpts were annotated. Some disagreement between the human annotators thus arose from this point of confusion. Although the annotators had previous knowledge of some of the novels, all annotations were made only with the context provided by each excerpt.

After the initial annotation phase, two annotators discussed each disagreement to create a set of consensus labels for the initial 96 excerpts. After the second round of annotations, an external adjudicator (native English speaker, post-secondary degree in text linguistics) decided on final annotations for all disagreements. Krippendorf's alpha was used after each annotation batch to

---

[3] *Salem's Lot*, *The Stand*, *The Dead Zone*, *Firestarter*, *Cujo*, and *The Gunslinger*

evaluate inter-annotator reliability. Finally, the additional 300 training samples were labeled by two annotators from the second batch of evaluation annotation. Disagreements were discussed and settled by the annotators.

3.3 Automated Annotation

We then evaluated the ability of two baselines, three BERT-based models, five Flan-T5 models, four Llama 3 models, and three models from OpenAI's GPT family to annotate for focalization. These models were chosen to represent a range of computational power, accessibility, and architectures. A detailed model list is available in Table 2.

The two baselines, a Naive Bayes classifier and logistic regression with word-count and TF-IDF features, were the least powerful but most accessible of the models we tested. These were included to determine whether focalization annotation is a trivial task that can be solved with only term frequencies. Similarly, DistilBERT, the two RoBERTA models, and the three smallest Flan-T5 models were evaluated to determine whether relatively lightweight models fine-tuned on a minimal dataset were sufficiently powerful to generate the annotations of interest. The two largest Flan-T5 models and the four Llama 3 models are larger and more powerful than the baselines, but are open source and accessible at no cost. Finally, the GPT models are the largest and the least accessible of the tested models as they require a per-token cost to query. These models assessed whether large, powerful models could identify focalization in a zero-shot setting.

We implemented the Naive Bayes and logistic regression classifiers and vectorizers using the `sklearn` package and trained them on an additional 300-sample training dataset. The DistilBERT model, RoBERTa models, and the three smallest Flan-T5 models were accessed and fine-tuned using HuggingFace. From the 300-sample training dataset, two excerpts labeled as zero focalization, eight labeled as external, and ten labeled as internal were included in the validation dataset, and the remaining 280 samples were used for training. Further fine-tuning parameters can be found in Appendix B. The two largest Flan-T5 models and the Llama 3 models were also accessed via HuggingFace. For the Llama models, "You are a helpful assistant." was passed as a generic system message. For each excerpt, the zero-shot prompt in Figure 1 was passed to the model with the excerpt of interest appended.

Finally, OpenAI's API was used to access the GPT models. Each time a GPT model was queried, we again passed "You are a helpful assistant." as a system message and the zero-shot prompt with the excerpt appended as a user message. The default parameter values were used for all but two parameters: nuclear sampling (`top_p`) and log-probabilities (`logprob`). `top_p` was set to 0.1, which forced the model to select tokens from only the top 10% of the probability mass. This was intended to optimize the model's performance for accuracy. `logprob` was set to `True` so the model output the log-probabilities for each token.

The log-probability values were then used to create a proxy 'confidence' metric for each annotation. Specifically, the log-probability for the first token of each annotation was converted into a probability p ($p = e^l$, where l is the log probability). Since no GPT model provided an annotation outside of the expected answer set (internal, external, or zero) and the first token of each of these modes is unique, the reported probability of the first token was interpreted as model confidence.

All baselines and models were used to annotate the evaluation dataset of 256 excerpts. The non-baseline models were asked to annotate the same texts three times, and the performance was averaged over all runs. Further, to test the highest-performing model's resistance to prompt variations, we created five alternate versions of the original prompt (Appendix C) and evaluated the model using each of these prompts. Finally, we used this same model to annotate all ≥ 50 word excerpts from each of the 16 novels in our subcorpus.

4 Annotation Accuracy and Agreement

4.1 Human Performance

Human annotators achieved mid-to-low inter-annotator reliability on this task (first round: α = 0.55, second round: α = 0.65), demonstrating its difficulty.[4] The annotators reported several primary challenges. First, they found it difficult to establish whether environmental information was communicated through a given character's perspective or through an external observer. For example, in the following quote, it is challenging to determine whether the use of "could see" is internally focalized through Trisha or externally through an observer:

> "Even in this channel, she was forced to clamber over one downed tree. It had fallen just recently, and 'fallen' was really the wrong word. Trisha could see more slash-marks in its bark, and (...) she could see how fresh and white the wood of the stump was."[5]

After discussion, the annotators agreed to label such instances as internally focalized, even if the information was externally observable, as long as there was some keyword indicating that it was perceived by a particular character (e.g. saw, heard). The annotators were also challenged by paragraphs where thoughts clearly belonging to a particular character were given as narration, as in the following quote:

> "Aye, as they said in these parts. If the boy had had the impertinence to begin an affair with the Mayor's gilly-in-waiting, and the incredible slyness to get away with it, what did that do to Jonas's picture of three In-World brats who could barely find their own behinds with both hands and a candle?"[6]

The annotators also agreed to label such examples as internally focalized. Overall, despite initial disagreements, the annotators reached consensus on labels for each excerpt in the evaluation dataset.

4.2 Model Performance

We evaluated the annotations produced by all computational baselines and LLMs against the consensus labels (Table 2). GPT-4o was the highest performing model, with an overall F1 score of 84.79%. All other models and baselines except for Llama 3.2 1b achieved considerably above-random performance, with F1 scores ranging from 55.25% (Flan-T5 small) to 80.82% (GPT-4-turbo). However, none performed on par with GPT-4o.

All models except Llama 3.2 1b achieved the highest F1 score for internal focalization, suggesting that it was the easiest mode to identify. In contrast, we find that the models nearly all performed the worst when annotating for zero focalization; in fact, of the seventeen tested models, nine failed to label any excerpts as zero-focalized. This may indicate the difficulty of identifying zero focalization, or may be a side-effect of class imbalance in the evaluation dataset.

Interestingly, the difference in F1 scores for the GPT models suggests that changes made between GPT-3.5 (first released March 15, 2022) and the GPT-4 models (first released March 14, 2023) significantly improved their ability to perform focalization annotation, and perhaps literary annotation more broadly. This difference in performance was reflected in the models' reported confidence values; GPT-3.5-turbo was, on average, less confident in its predictions than other

---

[4] Note that this is not uncommon for annotations of literary texts, which often exhibit comparable inter-annotator reliability scores, for example, in event type annotation [50].

[5] *The Girl Who Loved Tom Gordon* by Stephen King

[6] *Wizard and Glass* by Stephen King

Model	Internal F1	External F1	Zero F1	Overall Precision	Overall Recall	Overall F1
Logistic Regression*	81.23	62.86	0.0	69.96	73.83	71.84
Naive Bayes*	83.05	39.56	0.0	70.86	73.05	66.78
DistilBERT[†]*	85.16	61.07	0.0	73.94	77.60	74.01
RoBERTa Base[†]*	85.07	64.53	0.0	73.60	77.86	74.88
RoBERTa Large[†]*	86.06	71.65	0.0	75.54	79.82	77.47
Flan-T5 Small[†]*	69.56	30.59	0.0	55.61	59.64	55.25
Flan-T5 Base[†]*	82.34	23.28	0.0	73.49	71.09	61.92
Flan-T5 Large[†]*	86.47	67.31	0.0	75.97	79.82	76.58
Flan-T5 XL[†]	77.92	30.19	19.05	61.03	65.62	61.84
Flan-T5 XXL[†]	81.87	33.33	22.22	70.37	69.14	65.52
Llama 3.2 1b[†]	0.0	3.28	4.55	3.12	0.91	1.13
Llama 3.2 3b[†]	78.49	31.19	22.38	62.51	66.80	62.67
Llama 3.1 8b[†]	63.39	55.39	0.0	75.69	57.68	57.76
Llama 3.3 70b[†]	86.34	73.88	13.02	80.95	79.30	78.97
GPT-3.5-turbo[†]	75.02	64.19	29.82	76.56	68.49	69.63
GPT-4-turbo[†]	88.33	71.79	32.55	81.21	82.03	80.82
GPT-4o[†]	**88.73**	**84.78**	**36.16**	**86.63**	**84.64**	**84.79**

Table 2: Evaluations of model performance on focalization annotation. Overall scores are weighted by class size. Values reported for models marked with a dagger ([†]) are the averages over three runs, and those marked with an asterisk (*) were fine-tuned on 300 samples. All other models were prompted using the prompt in Figure 1. The values reported for Naive Bayes and logistic regression are the highest performing over several input feature variants; further results for these models are reported in Appendix D.

models (57.6% as compared to 97.2% (GPT-4-turbo) and 95.0% (GPT-4o)). This provides some evidence that the confidence values reflect model accuracy. An ANOVA test suggests we can reject the null-hypothesis, namely that the confidence values from all three models are drawn from the same distribution ($p < 0.001$[7]). The three GPT models also have low inter-annotator reliability ($\alpha = 0.47$).

However, GPT-4o demonstrated very high consistency across three model runs ($\alpha = 0.94$), with F1 scores only ranging from 84.1% to 84.7%. In this case, an ANOVA test does not refute the null hypothesis ($p \approx 0.90$), indicating no statistically significant variance between the confidence values produced by each model. This further suggests that GPT-4o can apply a consistent paradigm for annotating texts for focalization, achieving relatively high accuracy in alignment with human consensus annotations.

Furthermore, we found that GPT-4o was not particularly sensitive to variations in prompts. The model achieved a Krippendorf's alpha of 0.74 across all six prompt variants. The F1 scores for the variant annotations ranged from 79.8% (variant #3) to 84.5% (variant #5). Notably, Variant #5 provides the least explanation of all the tested prompts. Yet, it achieves nearly as high performance as the base prompt, which may indicate that GPT-4o had an understanding of focalization from pre-training. This resilience to prompt perturbations is notable given past research [1; 14; 18; 31; 52], and again suggests that GPT-4o is a reliable annotator of focalization.

Finally, we found that the confidence values produced by the GPT models corresponded to outside signals of the difficulty of annotating an excerpt. On average, texts that appeared to be

---

[7] p-value reported as 5.7×10^{-208}

	GPT 3.5-turbo		GPT 4-turbo		GPT 4o	
	Agree	Disagree	Agree	Disagree	Agree	Disagree
Humans	0.59 ($\pm$0.1)	0.54 ($\pm$0.2)	0.98 ($\pm$0.1)	0.96 ($\pm$0.1)	0.96 ($\pm$0.1)	0.91 ($\pm$0.1)
GPT Models	0.61 ($\pm$0.1)	0.52 ($\pm$0.1)	0.98 ($\pm$0.1)	0.96 ($\pm$0.1)	0.98 ($\pm$0.1)	0.91 ($\pm$0.1)
4o	0.58 ($\pm$0.1)	0.56 ($\pm$0.1)	0.98 ($\pm$0.1)	0.89 ($\pm$0.2)	0.96 ($\pm$0.1)	0.72 ($\pm$0.2)
4o (Prompts)	0.60 ($\pm$0.1)	0.53 ($\pm$0.1)	0.98 ($\pm$0.1)	0.95 ($\pm$0.1)	0.99 ($\pm$0.0)	0.85 ($\pm$0.2)

Table 3: The average of each model's confidence values when a subset of annotators (left column) agree or disagree. Standard deviations are given in parentheses. The differences in means are significant at $\alpha = 10^{-2}$ except for the confidence values of GPT 3.5-turbo when GPT 4o agrees and disagrees.

Novel	% Internal	% External	% Zero
The Girl Who Loved Tom Gordon	84.3	13.0	2.7
Dolores Claiborne	81.4	18.2	0.4
Cujo	68.5	25.3	6.3
Misery	66.5	30.5	3.1
The Green Mile	62.7	32.7	4.7
Insomnia	60.4	35.6	4.1
Firestarter	59.6	34.3	6.2
Desperation	59.5	37.6	2.9
Needful Things	58.9	36.2	4.9
The Stand	56.9	38.7	4.4
Wizard and Glass	55.5	38.5	6.1
The Waste Lands	55.2	38.4	6.3
The Gunslinger	52.4	38.4	9.3
The Eyes of the Dragon	48.7	23.6	27.7
The Dead Zone	47.9	46.4	5.7
Salem's Lot	45.7	49.6	4.7

Table 4: The percentage of paragraphs with ≥ 50 words from each novel that were annotated as internally, externally, or zero focalized.

more difficult to annotate received lower average confidence values (Table 3). This is true for texts where humans disagreed, the three GPT models disagreed, the GPT-4o runs disagreed, and GPT-4o with prompt perturbations disagreed. These differences are relatively small but are largely significant, suggesting that the confidence values correlate to some degree with the ambiguity of an annotation.

5 Focalization at Scale

To validate the usefulness of LLM annotations for focalization, we examine the distribution of focalization modes across the subcorpus of 16 Stephen King novels. Specifically, we examine whether the focalization annotations enable us to identify structural outliers in King's works and to show the association between focalization mode and the prominence of sensory descriptors.

5.1 Comparing Novel Structures

From the annotations created by GPT-4o, we find that the typical King novel is primarily internally focalized, with some externally and few zero focalized excerpts distributed regularly throughout its course. However, some novels deviate from this basic structure. In particular, two novels —

The Girl Who Loved Tom Gordon and *Dolores Claiborne* — have a much higher percentage of internally focalized paragraphs (Table 4). There are clear reasons why this is likely to be true for both texts. *The Girl Who Loved Tom Gordon* follows a young girl lost alone in the woods, and thus focalizes much of the narrative through her eyes. Additionally, *Dolores Claiborne* is written as an almost stream-of-consciousness narrative from the perspective of the titular character, and thus is again primarily internally focalized through that character.

Another outlier novel is *The Eyes of the Dragon*, which has a much higher percentage of zero focalized excerpts than any other novel (Table 4). This is likely because the novel has a fairy-tale-esque omniscient narrator, who often describes the perspectives of multiple or groups of characters at once, as in the following quote:

> "He had been allowed to marry late because he had met no woman who pleased his fancy, and because his mother, the great Dowager Queen of Delain, **had seemed immortal to Roland and to everyone else ⸺ and that included her**."[8]

Overall, we find that focalization annotations enable large-scale comparisons of novels, highlighting structural differences even within the works of a single author.

5.2   Links to Linguistic Features

Sense	Internal	External	Zero
Taste	*0.70	*-0.75	-0.06
Interoception	*0.54	*-0.65	0.06
Touch	*0.56	-0.34	-0.46
Smell	*0.51	-0.44	-0.21
Sound	-0.25	0.20	0.14
Sight	-0.21	0.34	-0.17

Table 5: Pearson's R correlations between sensorial information and focalization modes for each novel by King. All starred correlations have significance values $p < 0.05$.

Finally, we compared the annotations produced by GPT-4o to measures of other linguistic features. Scholars have previously hypothesized that internal focalization may be linked to a more *immersive* style, or one that attempts to pull readers into the reality of the text and make it "feel real" [3; 24; 25]. To explore the validity of this hypothesis, we examine a connection between focalization and sensory information, which we take as a proxy for immersion. While texts that are externally or zero focalized can contain sensory information, we seek to determine whether internally focalized texts are more likely to contain such sensory descriptors.

To test this hypothesis, we use the Lancaster Sensorimotor Lexicon [32], which contains sensorimotor strength values for ∼40,000 words across six perceptual axes: touch, hearing, smell, taste, vision, and interoception. For each of the 16 novels by King, we calculate the summed sensorimotor strength along each axis and standardize the value by the total number of words from the lexicon in the novel.[9] Finally, we examine the relationship between the mean sensorimotor strength along each axis for each novel and the percentage of excerpts from the novel that were labeled as internally or externally focalized (Table 5).

We find that there are relatively strong positive correlations between the percentage of internal focalized paragraphs in a novel and all perceptual axes except for sound and sight. There are

---

[8] *The Eyes of the Dragon* by Stephen King

[9] This encompasses nearly every word in the novel except for some function words like 'of' or 'is.'

particularly strong positive correlations between internal focalization and interoception (0.54) and taste (0.70), which may be less perceptible to an 'external' observer. Crucially, the opposite appears to be true for external focalization, which is negatively correlated with all senses except sound and sight. The correlations between the sensory values and the percentage of zero-focalized excerpts were generally weaker and not statistically significant. However, we found a notable negative correlation between zero focalization and touch. This may suggest that omniscient narrators focus less on characters' haptic sensory experiences, employ a language that is more reflective than descriptive [20], or may be an artefact of the skewed distribution of zero focalization values.

Overall, it appears that King is more likely to focus on the senses of smell, taste, touch, and interoception in internally focalized text and sound and sight in externally focalized text. This suggests that internally focalized texts may be more immersive, as they contain a greater variety of sensory descriptors, or at least that different forms of immersivity characterize King's uses of internal and external focalization. Whether this is true of focalization generally, rather than specific to King's writing, remains an open question.

6   Conclusion

In this paper, we demonstrated the usability of LLMs for a well-defined literary annotation task — namely, the presence of focalization in narrative discourse. We find that performance on this task varies across different model architectures and sizes, but that, broadly speaking, larger, prompted LLMs outperform smaller models fine-tuned on a minimal dataset. Perhaps unexpectedly, we find that, while the GPT family models are not terribly consistent with each other, the GPT-4o model is reliable across multiple runs with the same prompt and within prompt variants.

Overall, we conclude that zero-shot GPT-4o is an effective and reliable annotator of focalization and that other models, including Llama 3.3 70b, are promising, if not as reliable. In particular, the resilience of GPT-4o across prompt variations suggests that it consistently applies elements of literary theory and narratology. This indicates that LLMs hold great promise as annotators for a variety of literary tasks that require a subtle understanding of semantic and syntactic content, and that they will likely prove useful for even more abstract literary annotations.

References

[1]   Abraham, Louis, Arnal, Charles, and Marie, Antoine. "Prompt Selection Matters: Enhancing Text Annotations for Social Sciences with Large Language Models". 2024. DOI: `10.48550/arXiv.2407.10645`. arXiv: `2407.10645 [cs.CL]`.

[2]   Akimoto, Taisuke and Ogata, Takashi. "Experimental Development Of A Focalization Mechanism In An Integrated Narrative Generation System". In: *Journal of Artificial Intelligence and Soft Computing Research 5*, no. 3 (2015), pp. 177–188. DOI: `doi:10.1515/jaiscr-2015-0027`.

[3]   Allan, Rutger J, Jong, Irene JF de, and Jonge, Casper C de. "From Enargeia to Immersion: The Ancient Roots of a Modern Concept". In: *Style 51*, no. 1 (2017), pp. 34–51. DOI: `10.1353/sty.2017.0003`.

[4] Andringa, Els. "Effects of 'narrative distance' on readers' emotional involvement and response". In: *Poetics 23*, no. 6 (1996), pp. 431–452. DOI: `10.1016/0304-422X(95)00009-9`.

[5] Antoniak, Maria, Mire, Joel, Sap, Maarten, Ash, Elliott, and Piper, Andrew. "Where Do People Tell Stories Online? Story Detection Across Online Communities". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand, 2024, pp. 7104–7130. DOI: `10.18653/v1/2024.acl-long.383`.

[6] Bae, Byung-Chull, Cheong, Yun-Gyung, and Young, R. Michael. "Toward a computational model of focalization in narrative". In: *Proceedings of the 6th International Conference on Foundations of Digital Games*. Bordeaux, France, 2011, pp. 313–315. ISBN: 9781450308045. DOI: `10.1145/2159365.2159423`.

[7] Bal, Mieke and Boheemen, Christine van. *Narratology: Introduction to the Theory of Narrative*. Toronto, Canada: University of Toronto Press, 2009.

[8] Bamman, David, Chang, Kent K., Li, Lucy, and Zhou, Naitian. "On Classification with Large Language Models in Cultural Analytics". In: *Proceedings of the 5th Conference on Computational Humanities Research 2024 (CHR 2024)*. Aarhus, Denmark, 2024. DOI: `10.48550/arXiv.2410.12029`.

[9] Bjerring-Hansen, Jens, Al-Laith, Ali, Hershcovich, Daniel, Conroy, Alexander, and Rasmussen, Sebastian Ørtoft. "Literary Time Travel: Distinguishing Past and Contemporary Worlds in Danish and Norwegian Fiction". In: *Proceedings of the 5th Conference on Computational Humanities Research 2024 (CHR 2024)*. Aarhus, Denmark, 2024.

[10] Bruhns, Adrian and Köppe, Tilmann. "Internal Focalization and Seeing through a Character's Eyes". In: *Estetika: The European Journal of Aesthetics 61*, no. 2 (2024), pp. 126–141. DOI: `10.33134/eeja.364`.

[11] Brunner, Ann, Tu, Ngoc Duyen Tanja, Weimer, Lukas, and Jannidis, Fotis. "To BERT or not to BERT - Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation". In: *Proceedings of the Swiss Text Analytics Conference & Conference on Natural Language Processing 2020*. Zurich, Switzerland, 2020.

[12] Chatman, Seymour Benjamin. *Story and Discourse: Narrative Structure in Fiction and Film*. Ithaca, NY, USA: Cornell University Press, 1980. DOI: `10.1515/9781501741616`.

[13] Clasen, Mathias. "Why the World Is a Better Place with Stephen King in It: An Evolutionary Perspective". In: *Evolutionary Perspectives on Imaginative Culture*. New York City, NY, USA: Springer International Publishing, 2020, pp. 325–341. DOI: `10.1007/978-3-030-46190-4_17`.

[14] Coyne, Steven and Sakaguchi, Keisuke and Galvan-Sosa, Diana and Zock, Michael and Inui, Kentaro. "Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction". 2023. DOI: `10.48550/arXiv.2303.14342`. arXiv: `2303.14342 [cs.CL]`.

[15] Cranenburgh, Andreas van and Ketzan, Erik. "Stylometric Literariness Classification: the Case of Stephen King". In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 2021, pp. 189–197. DOI: `https://doi.org/10.18653/v1/2021.latechclfl-1.21`.

[16] Deleyto, Celestino. "Focalisation in Film Narrative". In: *Atlantis 13*, no. 1/2 (1991), pp. 159–177.

[17]   Emmott, Catherine. *Narrative Comprehension: A Discourse Perspective*. Oxford, United Kingdom: Clarendon Press, 1997. DOI: `10.1093/oso/9780198236498.001.0001`.

[18]   Gan, Chengguang and Mori, Tatsunori. "Sensitivity and Robustness of Large Language Models to Prompt Template in Japanese Text Classification Tasks". In: *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*. Hong Kong, China, 2023, pp. 1–11. DOI: `10.48550/arXiv.2305.08714`.

[19]   Genette, Gérard. *Narrative Discourse: An Essay in Method*. Ithaca, NY, USA: Cornell University Press, 1990.

[20]   Gittel, Benjamin, Barth, Florian, Dönicke, Tillmann, Gödeke, Luisa, Schomacker, Thorben, Varachkina, Hanna, Weimer, Anna Mareike, Holler, Anke, and Sporleder, Caroline. "Neither Telling nor Describing. Reflective Passages and Perceived Reflectiveness 1700-1945". Report. Darmstadt, Germany: Universitäts- und Landesbibliothek Darmstadt, 2024. DOI: `10.26083/tuprints-00027390`.

[21]   Herman, David. *Story Logic: Problems and Possibilities of Narrative*. Lincoln, NB, USA: University of Nebraska Press, 2002. DOI: `10.5860/choice.40-1361`.

[22]   Hicke, Rebecca and Mimno, David. "[Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs". In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*. St. Julians, Malta, 2024, pp. 270–277. DOI: `10.48550/arXiv.2401.17922`.

[23]   Hye-Knudsen, Marc, Kristensen-McLachlan, Ross Deans, and Clasen, Mathias. "How Stephen King Writes and Why: Language, immersion, emotion". In: *Orbis Litterarum 78*, no. 5 (2023), pp. 353–367. DOI: `10.1111/oli.12401`.

[24]   Jacobs, Arthur M and Lüdtke, Jana. "Immersion into narrative and poetic worlds". In: *Narrative absorption 27* (2017), p. 69. DOI: `10.1075/lal.27.05jac`.

[25]   Jumpertz, Jessica and Tary, Wiebeke. "An Empirical Study of Readers' Identification with a Narrator". In: *Anglistik 31*, no. 1 (2020), pp. 111–128. DOI: `10.33675/ANGL/2020/1/9`.

[26]   Kababgi, Daniel, Grisot, Giulia, Pennino, Federico, and Herrmann, Berenike. "Recognising non-named spatial entities in literary texts: a novel spatial entities classifier". In: *Proceedings of the 5th Computational Humanities Research Conference (CHR 2024)*. Paris, France, 2024, pp. 472–481.

[27]   Ketzan, Erik and Eve, Martin Paul. "The Anxiety of Prestige in Stephen King's Stylistics". In: *Journal of Computational Literary Studies 3*, no. 1 (2024), pp. 1–20. DOI: `https://doi.org/10.48694/jcls.3915`.

[28]   Kuzman, Taja, Mozetič, Igor, and Ljubešić, Nikola. "ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification". 2023. DOI: `10.48550/arXiv.2303.03953`. arXiv: 2303.03953 `[cs.CL]`.

[29]   Lijuan Chen, Xiaodong Xu and Lv, Hongling. "How literary text reading is influenced by narrative voice and focalization: evidence from eye movements". In: *Discourse Processes 60*, no. 10 (2023), pp. 675–694. DOI: `10.1080/0163853X.2023.2260247`.

[30]   Long, Hoyt and So, Richard Jean. "Turbulent Flow: A Computational Model of World Literature". In: *Modern Language Quarterly 77*, no. 3 (2016), pp. 345–367. DOI: `10.1215/00267929-3570656`.

[31] Lu, Yao, Bartolo, Max, Moore, Alastair, Riedel, Sebastian, and Stenetorp, Pontus. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland, 2022, pp. 8086–8098. DOI: `10.18653/v1/2022.acl-long.556`.

[32] Lynott, Dermot, Connell, Louise, Brysbaert, Marc, Brand, James, and Carney, James. "The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words". In: *Behavior research methods 52* (2020), pp. 1271–1291. DOI: `10.3758/s13428-019-01316-z`.

[33] Nelles, William. "Getting Focalization into Focus". In: *Poetics Today 11*, no. 2 (1990), pp. 365–382. DOI: `10.2307/1772622`.

[34] Niederhoff, Burkhard. "Focalization". In: *The Living Handbook of Narratology*. Berlin, Germany: De Gruyter, 2014.

[35] Niederhoff, Burkhard. "Perspective – Point of View". In: *The Living Handbook of Narratology*. Berlin, Germany: De Gruyter, 2014.

[36] Ørmen, Jacob and Gregersen, Andreas. "News as Narratives". In: *Oxford Research Encyclopedia of Communication*. Oxford, United Kingdom: Oxford University Press, 2019. DOI: `10.1093/acrefore/9780190228613.013.908`.

[37] Otmakhova, Yulia and Frermann, Lea. "Narrative Media Framing in Political Discourse". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria, 2025, pp. 9167–9196. DOI: `10.18653/v1/2025.findings-acl.477`.

[38] Pagel, Janis, Pichler, Axel, and Reiter, Nils. "Evaluating In-Context Learning for Computational Literary Studies: A Case Study Based on the Automatic Recognition of Knowledge Transfer in German Drama". In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*. St. Julians, Malta, 2024, pp. 1–10.

[39] Piper, Andrew, So, Richard Jean, and Bamman, David. "Narrative Theory for Computational Narrative Understanding". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic, 2021, pp. 298–311. DOI: `10.18653/v1/2021.emnlp-main.26`.

[40] Rebora, Simone, Lehmann, Marina, Heumann, Anne, Ding, Wei, and Lauer, Gerhard. "Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children's Literature". In: *Proceedings of the 4th Conference on Computational Humanities Research (CHR 2023)*. Paris, France, 2023, pp. 333–343.

[41] Rundquist, Eric. *Free Indirect Style in Modernism: Representations of Consciousness*. Linguistic Approaches to Literature. Amsterdam, Netherlands: John Benjamins Publishing Company, 2017. DOI: `10.1075/lal.29`.

[42] Sanders, José and Redeker, Gisela. "Linguistic perspective in short news stories". In: *Poetics 22*, no. 1-2 (1993), pp. 69–87. DOI: `10.1016/0304-422X(93)90021-8`.

[43] Sanford, A. J. and Emmott, Catherine. *Mind, Brain and Narrative*. Cambridge, United Kingdom: Cambridge University Press, 2012. DOI: `10.1017/CBO9781139084321`.

[44] Santana, Brenda, Campos, Ricardo, Amorim, Evelin, Jorge, Alípio, Silvano, Purificação, and Nunes, Sérgio. "A survey on narrative extraction from textual data". In: *Artificial Intelligence Review 56*, no. 8 (2023), pp. 8393–8435. DOI: `0.1007/s10462-022-10338-7`.

[45] Scholes, Robert, Phelan, James, and Kellogg, Robert. "Narrative Theory, 1966-2006: A Narrative". In: *The Nature of Narrative*. New York, NY: Oxford University Press, 2006, pp. 283–336. DOI: `10.1093/oso/9780195151756.003.0008`.

[46] Schubert, Christoph. "Forms and functions in speeches and hypertext frameworks: Narrative sequences in political discourse". In: *Narrative Revisited: Telling a story in the age of new media*, ed. by Christian R. Hoffmann. Amsterdam, Netherlands: John Benjamins Publishing Company, 2010, pp. 143–162. DOI: `10.1075/pbns.199.08sch`.

[47] Soni, Sandeep, Sihra, Amanpreet, Evans, Elizabeth F, Wilkens, Matthew, and Bamman, David. "Grounding Characters and Places in Narrative Text". In: *Proceedings of the 61st Annual Meeting Of The Association For Computational Linguistics*. Toronto, Canada, 2023. DOI: `10.18653/v1/2023.acl-long.655`.

[48] Stammbach, Dominik, Antoniak, Maria, and Ash, Elliott. "Heroes, Villains, and Victims, and GPT-3: Automated Extraction of Character Roles Without Training Data". In: *Proceedings of the 4th Workshop of Narrative Understanding (WNU 2022)*. Seattle, United States, 2022, pp. 47–56. DOI: `10.18653/v1/2022.wnu-1.6`.

[49] Thalken, Rosamond, Stiglitz, Edward, Mimno, David, and Wilkens, Matthew. "Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore, 2023, pp. 9252–9265. DOI: `10.18653/v1/2023.emnlp-main.575`.

[50] Vauth, Michael and Gius, Evelyn. "Event Annotations of Prose". In: *Journal of Open Humanities Data 8* (2022), p. 19. DOI: `10.5334/johd.83`.

[51] Walsh, Melanie, Preus, Anna, and Antoniak, Maria. "Sonnet or Not, Bot? Poetry Evaluation for Large Models and Datasets". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA, Nov. 2024, pp. 15568–15603. DOI: `10.18653/v1/2024.findings-emnlp.914`.

[52] Zhao, Zihao, Wallace, Eric, Feng, Shi, Klein, Dan, and Singh, Sameer. "Calibrate Before Use: Improving Few-Shot Performance of Language Models". In: *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*. Virtual, 2021, pp. 12697–12706. DOI: `10.48550/arXiv.2102.09690`.

A Corpus Contents

Title	Publication Date
Salem's Lot	1975
The Stand	1978
The Dead Zone	1979
Firestarter	1980
The Waste Lands	1981
Cujo	1982
The Gunslinger	1982
The Eyes of the Dragon	1987
Misery	1987
Needful Things	1991
Dolores Claiborne	1992
Insomnia	1994
Desperation	1996
The Green Mile	1997
Wizard and Glass	1997
The Girl Who Loved Tom Gordon	1999

B Fine-Tuning Parameters

Parameter	Value
Evaluation Strategy	epoch
Save Strategy	epoch
Learning Rate	2×10^{-5}
Weight Decay	0.01
# Train Epochs	5
Load Best Model at End	True

C Prompt Variants

Text in red indicates where changes from the original prompt (Figure 1) occurred.

Variant #1
Please state how the given novel excerpt is focalized, or the perspective through which narrative information is provided. There are three modes of focalization:

- INTERNAL: From the perspective of a particular character. Describes what the character knows, including their thoughts, behaviors, and emotions.

- EXTERNAL: From the perspective of an outside narrator. Describes characters' actions, behaviors, and settings. Communicates physically observable facts.

- ZERO: From every perspective. The narration has an omniscient point of view and can describe the thoughts, behaviors, emotions, and actions of any character.

Speech in quotation marks counts as external focalization. Only respond with one word representing the mode of focalization.

Variant #2

Please state the perspective through which narrative information is provided in the given novel excerpt. There are three modes ~~of focalization~~:

- INTERNAL: From the perspective of a particular character. Describes what the character knows, including their thoughts, behaviors, and emotions.

- EXTERNAL: From the perspective of an outside narrator. Describes characters' actions, behaviors, and settings. Communicates physically observable facts.

- ZERO: From every perspective. The narration has an omniscient point of view and can describe the thoughts, behaviors, emotions, and actions of any character.

Speech in quotation marks counts as external ~~focalization~~. Only respond with one word representing the mode ~~of focalization~~.

Variant #3

Please state how the given novel excerpt is focalized, or the perspective through which narrative information is provided. There are three modes of focalization:

- INTERNAL: ~~From the perspective of a particular character.~~ Describes what the character knows, including their thoughts, behaviors, and emotions.

- EXTERNAL: ~~From the perspective of an outside narrator.~~ Describes characters' actions, behaviors, and settings. Communicates physically observable facts.

- ZERO: ~~From every perspective.~~ The narration has an omniscient point of view and can describe the thoughts, behaviors, emotions, and actions of any character.

Speech in quotation marks counts as external focalization. Only respond with one word representing the mode of focalization.

Variant #4

Please state how the given novel excerpt is focalized, or the perspective through which narrative information is provided. There are three modes of focalization:

- INTERNAL: From the perspective of a particular character. ~~Describes what the character knows, including their thoughts, behaviors, and emotions.~~

- EXTERNAL: From the perspective of an outside narrator. ~~Describes characters actions, behaviors, and settings. Communicates physically observable facts.~~

- ZERO: From every perspective. ~~The narration has an omniscient point of view and can describe the thoughts, behaviors, emotions, and actions of any character.~~

Speech in quotation marks counts as external focalization. Only respond with one word representing the mode of focalization.

Variant #5

Please state how the given novel excerpt is focalized, or the perspective through which narrative information is provided. There are three modes of focalization: INTERNAL, EXTERNAL, ZERO

Speech in quotation marks counts as external focalization. Only respond with one word representing the mode of focalization.

D   Baseline Results

Features	Ngrams	Internal F1	External F1	Zero F1	Overall Precision	Recall	F1
			Logistic Regression				
Count	1	78.86	56.94	0.0	67.42	69.92	68.84
Count	1–2	81.23	**62.86**	0.0	69.96	73.83	**71.84**
Count	1–3	82.19	60.15	0.0	69.64	**74.22**	71.76
TF-IDF	1	**83.21**	46.46	0.0	**70.30**	73.83	68.75
TF-IDF	1–2	81.93	26.51	0.0	68.65	70.70	62.51
TF-IDF	1–3	81.15	17.72	0.0	65.57	69.14	59.61
			Naive Bayes				
Count	1	**83.05**	**39.56**	0.0	70.86	**73.05**	**66.78**
Count	1–2	82.58	25.32	0.0	**74.48**	71.48	62.63
Count	1–3	81.99	18.42	0.0	73.90	70.31	60.37
TF-IDF	1	80.65	0.0	0.0	45.67	67.58	54.50
TF-IDF	1–2	80.65	0.0	0.0	45.67	67.58	54.50
TF-IDF	1–3	80.65	0.0	0.0	45.67	67.58	54.50

Table 6: Evaluations of model performance on focalization annotation for all Naive Bayes and Logistic Regression models. The highest value in each column is bolded.