Do Language Models Agree with Humans Perceptions of Suspense in Stories?

Glenn Matlin[*], **Devin Zhang**[◇], **Rodrigo Loza**[◇], **Diana M. Popescu**[◇],
Jacqueline Isbell, Chandreyi Chakraborty, Mark Riedl
School of Interactive Computing, College of Computing
Georgia Institute of Technology, USA

Correspondence: glenn@gatech.edu

Abstract

Suspense is an affective response to narrative text that is believed to involve complex cognitive processes in humans. Several psychological models have been developed to describe this phenomenon and the circumstances under which text might trigger it. We replicate four seminal psychological studies of human perceptions of suspense, substituting human responses with those of different open-weight and closed-source LMs. We conclude that while LMs can distinguish whether a text is intended to induce suspense in people, LMs cannot accurately estimate the relative amount of suspense within a text sequence as compared to human judgments, nor can LMs properly capture the human perception for the rise and fall of suspense across multiple text segments. We examine the functional limits of LMs suspense understanding capabilities by adversarially permuting the story text to probe what causes the perception of suspense to diverge for both human and LMs. We conclude that, while LMs can superficially identify and track certain facets of suspense, they do not process suspense in the same way as human readers.

1 Introduction

Suspense, a multifaceted emotional phenomenon of the human mind, has been extensively studied across psychology (Gerrig, 1994; Madrigal et al., 2011), cognitive science (Colby et al., 1989; Hoeken & van Vliet, 2000) literary analysis (Brewer & Lichtenstein, 1982; Brewer, 1996), and film studies (Branigan, 1992; Carroll, 1984; Vorderer et al., 1996). In storytelling—a fundamental medium for communication, entertainment, education, conveying intricate ideas, building rapport, and social commentary—suspense plays a pivotal role by engaging audiences in complex emotional states (Gerrig, 1994). Stories capable of generating suspense can heighten a reader's emotional investment, making them more likely to connect deeply with the content (Doust & Piwek, 2017; O'Neill & Riedl, 2014). Whether in games (Li et al., 2021), news (Kaspar et al., 2016), sports (Peterson & Raney, 2008), or business (Hsieh, 2021; Guidry, 2005; Moulard et al., 2012; Palich & Bagby, 1995), the strategic evocation of suspense shapes individuals' emotional experiences and decision-making processes. By amplifying tension and creating unresolved questions, suspense compels readers or viewers to continuously seek resolution. Suspense is believed to be an affective response to narrative content, written, visual, or otherwise, that involves complex cognitive phenomena such as empathy, perceptions of uncertainty, and anticipation (Gerrig, 1994).

Suspense has also been a topic of study in Artificial Intelligence (AI), including the detection of suspense in narrative texts (Fendt & Young, 2017; Delatorre et al., 2020; 2021; O'Neill & Riedl, 2011; 2014; Doust & Piwek, 2017) and the generation of texts that evoke suspense in human readers (Fendt & Young, 2017; Delatorre et al., 2020; 2021; Xie & Riedl, 2024) Furthermore, AI models that can detect sophisticated narrative devices such as suspense

[*]:Corresponding author, [◇]: Equal contribution, `https://github.com/eilab-gt/SuspensePerception`

may be used to develop intelligent human creativity support tools. While writing support is often approached from a generative standpoint, constructive critique can also support human creators (Lin et al., 2023). Creative writing therefore serves as a useful benchmark for testing the capabilities of LMs (Paech, 2023) and AI agents utilizing LMs (Gooding et al., 2025) In this paper we concern ourselves with the **detection** and **rating** of suspense in narrative text. Early attempts at computationally modeling suspense relied on symbolic representations (Cheong & Young, 2015; O'Neill & Riedl, 2011; 2014; Doust & Piwek, 2017). However, these techniques required manual knowledge engineering to convert stories into symbolic representations for analysis. Language Models (LMs) based on the Transformer architecture (Vaswani et al., 2017) have shown remarkable capabilities in tasks such as sentiment analysis, summarization, and question answering. The question of whether LMs can make judgments of story suspense that correspond to human perceptions is not necessarily straightforward. Suspense is a *complex affective phenomena* potentially involving empathy, perceptions of uncertainty, and prediction of outcomes. This complexity has meant that LMs are not particularly adept at generating suspenseful narratives (Xie & Riedl, 2024). Generation is not the same as detection, which leads to our overall research question: how does the LM rating of suspense in written narrative texts relate to human perceptions of suspense in the same stories?

Our study investigates the ability of current LMs to detect and measure suspense in textual passages. Specifically, we pose four research questions aimed at probing how closely LMs can mirror human suspense perception: **(1)** Can LMs distinguish whether a text sequence is intended to generate suspense? **(2)** Do LMs accurately estimate the *relative* amount of suspense within a text sequence compared to human judgments? **(3)** Can LMs identify the key moments when story suspense *rises and falls* across multiple text segments, in alignment with human ratings? **(4)** Does adversarially permuting story text cause a divergence between human and LM perceptions of the text? We conduct a *machine psychology* study by replicating four notable psychology studies of human perceptions of suspense by swapping LMs for human participants. We design our experiment to follow the best practices in machine psychology (Löhn et al., 2024). We investigate multiple families of LMs at a variety of parameter sizes, both open and closed source. We compare LM results to the data from the human participant studies.

Our findings lead us to conclude that while LMs can superficially identify and even track certain facets of suspense, they do not process or represent suspense in the same way human readers do. LMs struggle to quantify suspense levels on a continuous scale, and the machine's response often fails to accurately align with human judgments. LMs stumble even when identifying the most critical moments of rising and falling action across a narrative. Our experiments with adversarial or context-disrupting permutations show that LMs do not significantly alter their generated answers, even when the attacked text is significantly altered to increase/decrease suspense. Overall, we conclude that the surveyed LMs do "mimic" suspense judgments to a certain extent, but the manner in which they do so is sufficiently brittle—and often context-insensitive—that the measures made by these LMs diverge from the dynamic, human experience of unfolding tension.

2 Related Work

Understanding whether LMs can detect suspense falls within the domain of *machine psychology* (Hagendorff, 2023), which applies psychological experiments, originally designed to study human cognition and behavior, to investigate computational models of intelligence and behavior. Several notable studies have directly examined and quantified human perceptions of suspense in stories, yielding alternative explanations for how the feeling of suspense arises from reading. Gerrig & Bernardo (1994) models suspense as a function of an individual's assessment of the possible solutions within a given problem and the complexity of transitioning from an initial state to a desired outcome. Wilmot & Keller (2020) explores these transitional states, proposing that a hierarchical language model leveraging uncertainty reduction can closely match human annotations of suspense in short stories. Lehne & Koelsch (2015) proposes a cognitive model of suspense that conceptualizes it as an affective state driven primarily by uncertainty and anticipation about future narrative outcomes.

Brewer & Ohtsuka (1988) developed Structural-Affect Theory, positing that suspense arises specifically when an initiating narrative event generates emotional uncertainty by delaying the resolution or outcome. Delatorre et al. (2018) measures the effects of uncertainty on suspense perception by providing one group of participants the ending in advance. Brewer & Ohtsuka (1988) tests the Structural-Affect Theory, hypothesizing that the structural feature of an initiating event and the inclusion of its outcome produce suspense. Haider et al. (2020) tests language model BERT on emotion classification, including suspense, using their annotated dataset on poetry. They found suspense was among the emotions the model had the most difficulty predicting. Bentz et al. (2024) produces a novel experiment to rate the suspense arc of a narrative in real time. As participants read a story, they drew a line that correlated to their perceived suspense levels.

From these and other papers, we drew our datasets based on the requirement that the papers could be replicated under the machine psychology experimental conditions suggested by (Löhn et al., 2024) and provided quantifiable data in the form of human-annotated ratings of textual suspense on a Likert scale. From this criteria we arrived at our specific experiment dataset using results from Gerrig & Bernardo (1994), Brewer & Ohtsuka (1988), Lehne & Koelsch (2015), and Delatorre et al. (2018).

Sentiment analysis (Liu, 2012; Wankhade et al., 2022) and emotional analysis (Plaza-del Arco et al., 2024) attempt to classify text in terms of whether the author is attempting to convey a sentiment or whether a character is in a particular emotional state (Zhu et al., 2011). While suspense is also an affective response, suspense differs from the above in that suspense is the emotion that the *reader* will experience upon consuming narrative content.

Early work aimed at computationally model and detect suspense used symbolic logical systems (Cheong & Young, 2015; O'Neill & Riedl, 2011; 2014; Doust & Piwek, 2017). Steg et al. (2022) trained machine learning models to classify suspense in texts using reader annotations. They collected annotations on suspense, curiosity, and surprise, as well as human ratings from Piper et al. (2021), who assessed situatedness, event sequencing, and world-making. Wilmot & Keller (2020) analyzed suspense in short stories by computing vector embeddings and measuring uncertainty reduction through a forward-looking probability distribution of possible narrative paths. This method proved to be a strong predictor of human-perceived suspense, as it represented when resolution appeared imminent. Our work directly leverages LLMs for suspense detection rather than traditional NLP models, benefiting from their ability to capture long-range dependencies in text and training on large corpora.

3 Experimental Setup

In this section, we detail our methodology for evaluating Large Language Models (LMs) on the task of suspense detection. By adapting classic psychology experiments to a modern AI context, we explore whether models can mirror the same rise, peak, and resolution of suspense that human readers experience. To conduct a *machine psychology* experiment, we follow the recommendations and requirements developed in Hagendorff (2023) for how to reduce bias in machine psychology studies and ensure a fair comparison across models and with humans.

3.1 Dataset and Sources

We replicate four foundational psychology studies to capture the diverse ways in which suspense manifests. These works vary in text length, genre, and annotation style, allowing us to test whether LMs can generalize across narrative structures and psychological instruments. Each employed rigorous methodologies to obtain human suspense ratings. Selection criteria prioritized datasets that offered demographic diversity and variation in text segment length, ensuring a broader representation of suspense perception. We specifically selected the studies that utilized a Likert scale. This choice enables a quantifiable comparison between human ratings and model predictions.

Gerrig & Bernardo (1994) used excerpts from a James Bond novel, originally rated by participants at Yale University and the University of the Philippines on two dimensions: *likelihood of escape* and *suspense*. Each excerpt was assigned a numeric rating on a 1–7 scale, reflecting the intensity of participants' reactions. We maintain the experimental conditions (e.g., presence or absence of a key item) and replicate these scenarios for the LM prompts.

Brewer & Ohtsuka (1988) introduced a structural-affect theory of stories, using American and Hungarian short stories. Participants rated suspense at five key plot segments for each story. This segmentation is ideal for our multi-step model prompts, as it requires the LM to *incrementally* track how suspense evolves as the narrative unfolds.

Delatorre et al. (2018) investigated how the valence of the ending (good/bad), the revelation of the ending (revealed/unrevealed), and writing style (journalistic/novelistic) affect suspense in an English-language short story. We use the original twelve segments as published and preserve the original rating scale (1–9) The authors intentionally designed the story narrative used in the original experiment to follow the classical pyramidal structure proposed by Freytag (1894), explicitly incorporating stages of exposition, rising action, climax, falling action, and denouement (see Figure 5). The rising action deliberately introduces uncertainty while the falling action explicitly resolves this suspense.

Lehne & Koelsch (2015) measured segment-level suspense in a shortened German-language version of *'The Sandman'* (1816; German: *Der Sandmann*) by E. T. A. Hoffmann. We used the publisher's official English translation divided into 65 segments matching the original German experiment. The human participants rated each segment in real time. The question is whether LMs similarly track the continuous "ebb and flow" of suspense across a longer narrative, aligning with human data.

3.2 Replicating the Studies with LMs

We replicate each study's instructions nearly verbatim, ensuring that the LM produces a numeric value within the specified range used in each original human study. If the original research collected multiple ratings along different dimensions (e.g., "likelihood of escape" and "suspense"), we prompt the model to provide both. Brewer asks some additional questions along several dimensions beyond suspense (i.e., rating the segment's ironic connotation and its relation to the suspense); we did not replicate these questions. Gerrig's participants were provided with entire passages of text before answering questions. Brewer, Delatorre, and Lehne conducted their studies using a text sample that spanned multiple segments or the story. The resulting labels from human annotators from each study were provided at an aggregate level, limiting our ability to analyze inter-rater agreement or similar annotation quality metrics. To replicate these studies, the machine receives each segment sequentially, which prevents LMs from attending to elements that will appear later in the story than each particular segment being asked about. The full experiment and story are conducted in a multi-message conversational format. During this conversation, we provide the instructions (e.g., "Rate how suspenseful you find this passage on a 1–7 scale") to every segment, staying in line with the requirements of a machine psychology experiment. We append the message and response from the first segment inquiry to the front of the second prompt and segment, continuing in this manner until the story is finished and the experiment is completed.

3.3 LM Selection and Configurations

We evaluate multiple state-of-the-art LMs that were available and widely used contemporaneously with our study. We selected these language models based on their high performance on numerous common language modeling leaderboards (Liang et al., 2022; Wang et al., 2024). We select LMs to offer as much diversity as possible by selecting LMs that differ in parameter count, training corpus, and availability (closed vs. open source). To holistically evaluate suspense, we focus our evaluation on foundational instruction-tuned language models (LMs) and do not fine-tune models specifically for the suspense detection task. **open-weight models** include: Llama2 7B, Llama3 8B, Llama3 70B, Gemma2 9B, Gemma2 27B, DeepSeek-V3, Qwen2 72B, Mistral 7B, Mixtral 8x7B, and WizardLM2 8x22B. To minimize

variability between samples, we set the *temperature* to 0 for models where we can control the randomness of sampling. We repeat each experiment three times and average the outputs, while acknowledging the residual randomness of the model that we cannot directly seed.

3.4 Evaluation Metrics

We aim to evaluate the extent to which LMs accurately capture the presence, relative magnitude, and trajectory of suspense. All the experiment data from Gerrig & Bernardo (1994), Brewer & Ohtsuka (1988), Delatorre et al. (2018), and Lehne & Koelsch (2015) used numerical ratings with a middle neutral point on a scale of either 1–7 or 1–9. We measure the **correlational error** using standard metrics between the model's generated rating value and the mean human rating value. Mean Absolute Error (MAE) measures the distance between the LM predictions and the human average.

Recognizing that an LM may have a different understanding of numerical scales than humans and demonstrate specific biases favoring certain numbers or scale (Shaki et al., 2023; Shah et al., 2023; Shaikh et al., 2024; Malberg et al., 2024), we present results when we relax the requirement for the LM to directly predict the magnitude of the numeric value. We analyze performance when performing binarization the machine's generated values on the Likert scale into a binary label representing the level of "high" or "low" suspense. Using a binary classification, we provide analysis regarding whether the generated values by these LM would label the segments as having "high" or "low" suspense, mimicking how human participants suspense would be anchored to the midpoint separating the upper or lower halves of the Likert scale.

We also look at the **directionality** of suspense ratings—whether suspense is increasing or decreasing from one segment of story to the next. We compute the above metrics on a per-segment level, we can also see where suspense perceptions are rising or falling and whether the LM's ratings are also rising or falling between the same pair of segments.

Each model is run three times on each story segment, after which we take the mean rating per segment. We then calculate metrics using averaged outputs. This repetition partially offsets unknown sampling variance for closed-source models and ensures consistent reporting.

4 Results

The results for rating agreement for Gerrig are in Figure 6, which is provided as a heatmap. Each column is a story treatment—Gerrig gives different stories with variations to increase or reduce suspense. Each row is a different model, and the bottom row has the average human ratings. Yellow indicates close match between model and human ratings, and blue indicates a large discrepancy (too high or too low). The heatmap is designed to allow for visual inspection of the general trends of models and stories. Appendix C gives full-size and charts the details of the story variations. Generally, the LMs vary significantly from human ratings, either reporting too much suspense or too little. Table 1 shows agreement when we binarize the human data, with any rating above the middle rating indicating presence of suspense. The LMs are generally in agreement with humans when we binarize the data. We note that the story variations decrease suspense but never below the middle point, and the LMs report all stories suspenseful. We also note that the Gerrig experiment uses James Bond stories, which are part of all LM pre-training data.

Figures 7a, 7b, and 7c show results for Brewer, Delatorre, and Lehne, respectively (Brewer had multiple stories and we only show the results for one story here; others can be found in Appendix C). In these experiments, storys are broken into segments. Generally speaking, the results are a mix of close matches and misses. As before, full size heatmaps are in Appendix C. Larger models tend to produce more accurate matches, but not uniformly so. Table 2 show matches on binarized data where we see agreement at 78% or above.

Because Delatorre and Lehne have stories broken into segments and have sufficient length, we are able to conduct directionality analysis wherein we mark agreement if the segment-to-segment ratings are rising or falling in accordance with rising or falling human ratings.

See Figure 8 (top). We see that model ratings trend in the same direction as human ratings, indicating that LMs can accurately assess the coarser notion of rising and falling suspense. There are vertical bands of darker colors in the heatmaps that warrant further analysis. These bands correlate with a switch from rising suspense to falling suspense, or vice-versa, according to human ratings. When we isolate and analyze these inflection points where rising changes to falling (Figure 8 bottom), we see that LMs are substantially less reliable. However, when we look at binarized scales (suspense present/not-present), agreement does not noticeably drop (see Appendix C).

Most significantly, LMs failed to detect suspense where humans succeeded for passages that increase suspense in the context of the story while being not inherently suspenseful on their own. Due to the fact that we fed the models text in chunks for Brewer, Delatorre, and Lehne with the repeated prompt, we considered that the LMs may be answering the prompt without considering how the text relates to the storyline as a whole and reran the LMs so the rating prompt "Rate the following paragraph for its suspensefulness on a 9-point scale, where 1 is 'not suspenseful' and 9 is 'very suspenseful' " included "in relation to the storyline so far." at the end. However, we found that this did not significantly positively or negatively impact the results of the LMs.

Model	Acc. ↑	F1 ↑	MSE ↓
Qwen2 72B	1.00 ± 0.00	1.00 ± 0.00	13.38 ± 4.33
DeepSeek V3	1.00 ± 0.00	1.00 ± 0.00	13.38 ± 4.33
Gemma 2 27b	1.00 ± 0.00	1.00 ± 0.00	12.02 ± 4.82
Llama 2 7b	1.00 ± 0.00	1.00 ± 0.00	13.38 ± 4.33
Llama 3 70b	1.00 ± 0.00	1.00 ± 0.00	13.38 ± 4.33
Llama 3 8b	1.00 ± 0.00	1.00 ± 0.00	26.15 ± 7.15
WizardLM 2 8x22B	1.00 ± 0.00	1.00 ± 0.00	13.38 ± 4.33
Mistral 7B	1.00 ± 0.00	1.00 ± 0.00	7.84 ± 5.48
Mixtral 8x7B	1.00 ± 0.00	1.00 ± 0.00	8.87 ± 6.41
Average	1.00 ± 0.00	1.00 ± 0.00	13.18 ± 5.21

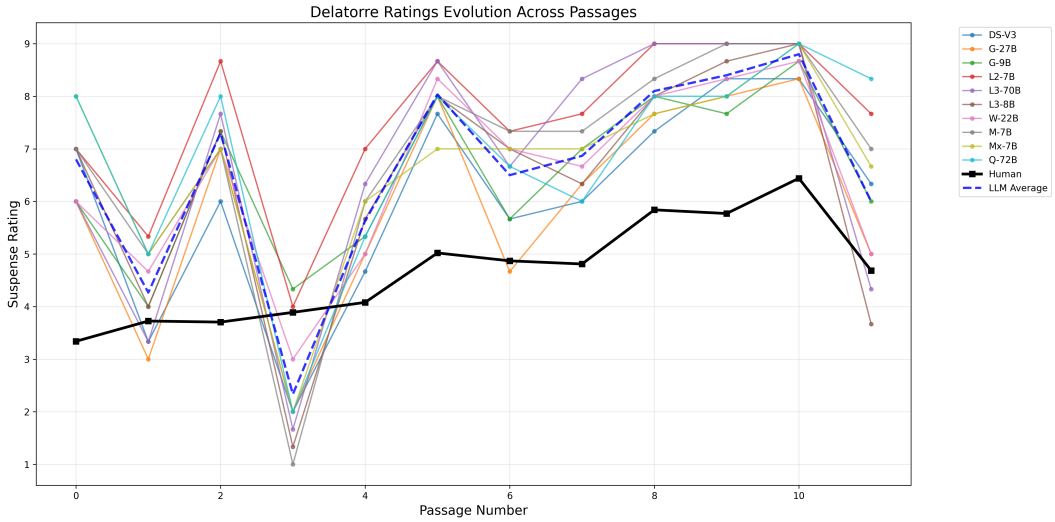
Table 1: Gerrig results on binary yes/no scale.

	Brewer			Delatorre			Lehne		
Model	Acc. ↑	F1 ↑	MSE ↓	Acc. ↑	F1 ↑	MSE ↓	Acc. ↑	F1 ↑	MSE ↓
Qwen2 72B	1.00 ± 0.00	1.00 ± 0.00	4.61 ± 2.09	0.68 ± 0.10	0.77 ± 0.08	8.81 ± 1.68	0.80 ± 0.00	0.88 ± 0.00	4.53 ± 0.00
DeepSeek V3	0.93 ± 0.15	0.96 ± 0.09	8.08 ± 4.54	0.69 ± 0.12	0.79 ± 0.09	6.54 ± 0.68	0.86 ± 0.00	0.91 ± 0.00	4.17 ± 0.00
Gemma 2 27b	0.80 ± 0.27	0.87 ± 0.18	5.31 ± 2.51	0.78 ± 0.06	0.85 ± 0.04	5.13 ± 0.79	0.86 ± 0.00	0.92 ± 0.00	5.35 ± 0.00
Gemma 2 9b	0.80 ± 0.33	0.85 ± 0.26	3.84 ± 2.77	0.68 ± 0.08	0.80 ± 0.05	7.47 ± 2.11	0.83 ± 0.00	0.90 ± 0.00	5.90 ± 0.00
Llama 2 7b	0.19 ± 0.38	0.21 ± 0.43	18.47 ± 14.97	0.68 ± 0.05	0.81 ± 0.04	12.19 ± 1.51	0.83 ± 0.00	0.91 ± 0.00	12.95 ± 0.00
Llama 3 70b	0.72 ± 0.31	0.80 ± 0.25	4.79 ± 3.74	0.72 ± 0.11	0.79 ± 0.11	12.22 ± 1.91	0.77 ± 0.00	0.84 ± 0.00	7.87 ± 0.00
Llama 3 8b	0.92 ± 0.18	0.95 ± 0.11	1.95 ± 0.83	0.65 ± 0.12	0.76 ± 0.11	9.92 ± 1.00	0.72 ± 0.00	0.81 ± 0.00	8.83 ± 0.00
WizardLM 2	0.80 ± 0.45	0.80 ± 0.45	6.88 ± 5.97	0.70 ± 0.09	0.80 ± 0.07	7.85 ± 3.51	0.83 ± 0.00	0.89 ± 0.00	4.09 ± 0.00
Mistral 7B	0.76 ± 0.36	0.82 ± 0.29	1.52 ± 0.58	0.68 ± 0.08	0.79 ± 0.07	8.27 ± 2.13	0.85 ± 0.00	0.91 ± 0.00	10.58 ± 0.00
Mixtral 8x7B	0.45 ± 0.37	0.55 ± 0.37	9.20 ± 10.61	0.69 ± 0.07	0.80 ± 0.05	10.22 ± 2.01	0.71 ± 0.00	0.79 ± 0.00	5.88 ± 0.00
Average	0.74 ± 0.28	0.78 ± 0.24	6.46 ± 4.86	0.69 ± 0.09	0.80 ± 0.07	8.86 ± 1.73	0.81 ± 0.00	0.88 ± 0.00	7.02 ± 0.00

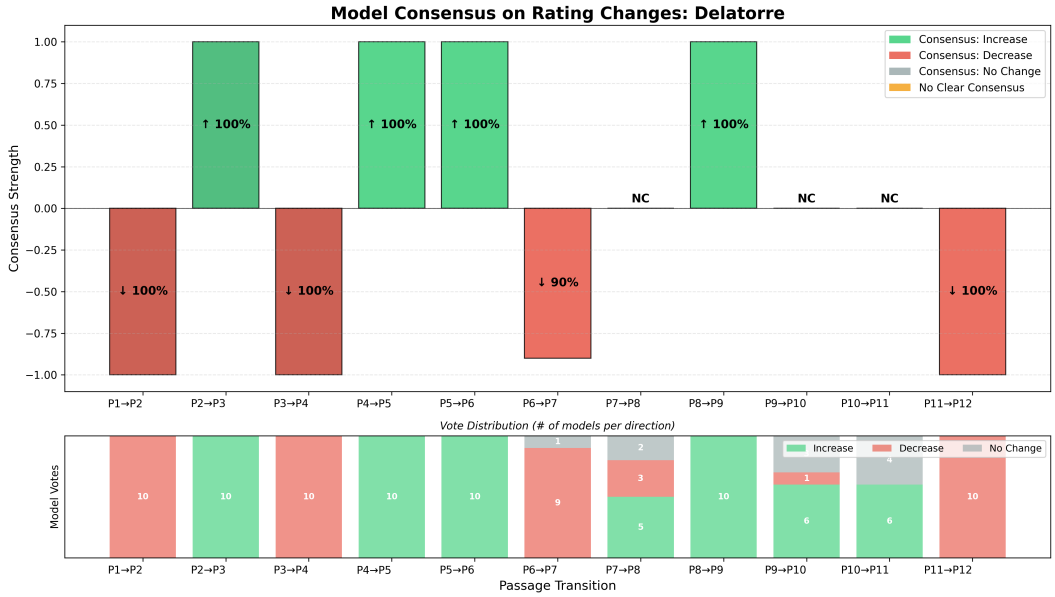
Table 2: Brewer, Delatorre, and Lehne results on binary yes/no scale.

5 Adversarial Story Experiments

According to psychological models of suspense, complex cognitive processes–such as reasoning about the probability and uncertainty of potential story outcomes–underlie the experience of suspense in humans (Gerrig & Bernardo, 1994; Brewer & Ohtsuka, 1988; Delatorre et al., 2018; Lehne & Koelsch, 2015). Brewer's Structural-Affect Theory explicitly connects suspense to emotional uncertainty resulting from delayed narrative outcomes (Brewer & Ohtsuka, 1988). Lehne's model emphasizes uncertainty and anticipation about the narrative's future outcomes as key drivers of suspense (Lehne & Koelsch, 2015). Similarly, Delatorre et al. (2018) experimentally validated suspense as dependent on outcome uncertainty by demonstrating significant suspense reduction when narrative endings were revealed in advance. Given that suspense fundamentally relies on such complex cognitive mechanisms, the question arises: why should an LM, devoid of human cognitive faculties, be able to detect and rate suspense in narrative texts? Are LMs simply triggering on keywords or have they internalized textual patterns and narrative tropes associated with suspense? Alternatively, might LMs exhibit a form of "reasoning" about potential outcomes,



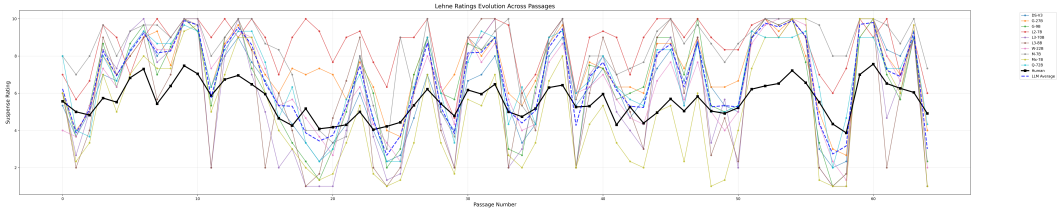
(a) Delatorre Visualization of Model Results.



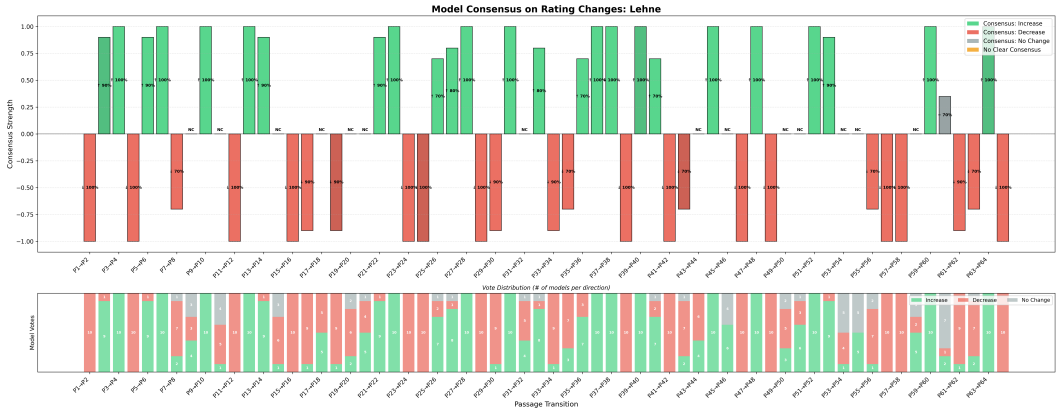
(b) Delatorre Visualization of Model Consensus.

as activation patterns propagate through successive layers of their network architectures? Lastly, could their performance reflect exposure to the very texts or types of studies typically used in psychological research?

To investigate how LMs process stories for suspense detection, we conduct 11 types of adversarial permutations on stories from Gerrig & Bernardo (1994) and Delatorre et al. (2018). These adversarial attacks target surface forms (e.g., word-order manipulations) and shallow semantic levels (e.g., name substitutions), aiming to increase the reader's cognitive load. Forecasting potential outcomes (a key element of many suspense theories) is already a high-cognitive-load process (Graesser et al., 1994). Increasing demands on cognitive resources has been shown to interfere with inferential comprehension and prediction (**?**), suggesting that when readers must devote attentional capacity to resolving text disruptions, fewer resources remain for generating or updating outcome forecasts. Consequently, if suspense depends upon these predictive inferences, deliberately raising the cognitive load could diminish suspense perceptions. Our adversarial permutations are chosen to be implementable at scale without the need for the careful plot treatment designs such as was



(a) Lehne directionality.



(b) Lehne inflection points.

done in the original Gerrig study. These adversarial tests serve as a stress test for both human readers and LMs to probe whether and how they engage in such forecasting under heightened cognitive demands.

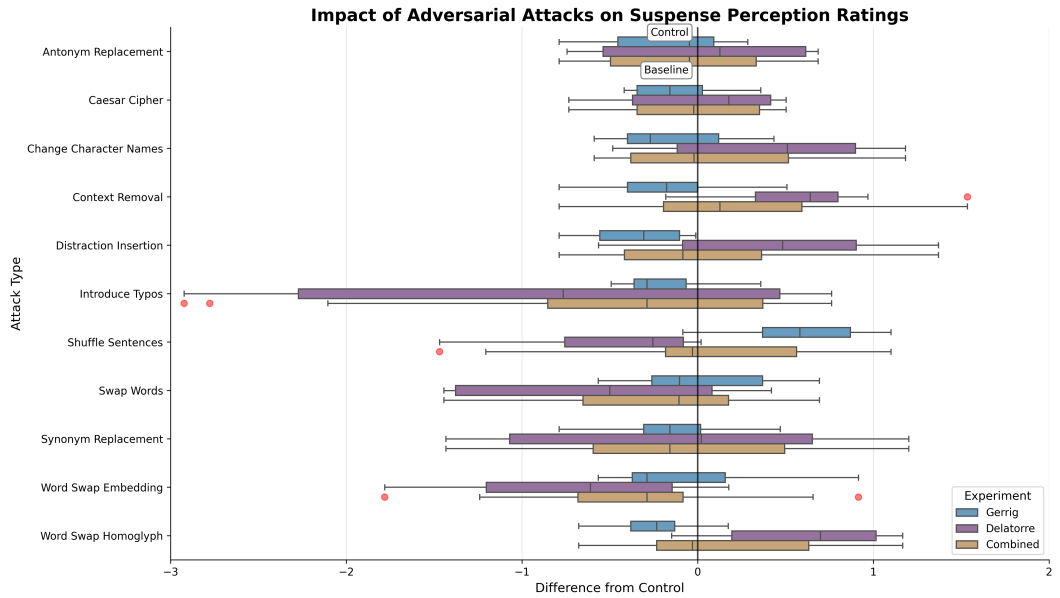


Figure 3: Summary of adversarial results on Gerrig and Delatorre.

Our adversarial permutations are as follows:

1. **Word Swapping** Pairs of words within each passage are swapped. This is destructive to readability from a human perspective. Depending on the number of swaps, the text can become nonsensical, making it difficult for humans to understand.

2. **Antonym Substitution** Available words in each sentence are replaced by their antonyms according to the WordNet lexical database, with probability of 80%. Each passage is limited to having between 1 and 50 antonyms.
3. **Synonym Substitution** Between one and ten words per passage are replaced with synonyms, with the same approach and parameters used for Antonym Substitution.
4. **Caesar Ciphering** Text is ciphered using a Caesar cipher with a step size of 3. This method is idealized as being highly disruptive to a human reader. For this attack, the prompt is modified to include information about the cipher and the step size.
5. **Homoglyph Substitution** Words are replaced with visually similar homoglyphs. This attack is idealized as being barely perceivable to humans.
6. **Distraction Phrase Insertion** A series of preset distraction phrases are inserted according to a "budget" of one distraction every five sentences. This phrase is unrelated to the text but is semantically meaningful. We do not anticipate that the phrase would be factored in the human understanding of suspense, and do not include relevant suspense trigger words.
7. **Sentence Shuffling** Sentences are globally shuffled, altering the semantic content of the text.
8. **Embedding-Based Word Substitution** Words are swapped based on their nearest neighbors in GloVe (Pennington et al., 2014) embeddings.
9. **Character Name Changes** Character names were changed.
10. **Context Removal** Sentences with sentiment scores below a threshold (0.5 compound sentiment) were removed, measured via VADER (Hutto & Gilbert, 2014) sentiment analysis model.
11. **Typo insertion** Up to 10 characters per sentence are replaced with random characters.

We run the same experimental setup from §3: EXPERIMENTAL-SETUP but with our permuted story segments. For each of the adversarial attacks, we expect human ratings of suspense to **drop** because the texts are made unreadable or nonsensical in different ways to increase the cognitive load of reading. Thus we look for situations where the LM's ratings **remain the same**, or **rise** as evidence that LMs are using very different processes to generate ratings.

Figure 3 shows the changes in LM ratings when each adversarial permutation is applied across all story segments from Gerrig & Bernardo (1994) and Delatorre et al. (2018). For most attacks, the average LM response remains unchanged, with less than a point of rating difference in either direction. Some attacks cause a *rise* in suspense in the Delatorre et al. (2018) story segments, especially when character names are changed or distraction sentences are inserted. Because models should be able to operate on all types of stories, the results on the combined Gerrig & Bernardo (1994) and Delatorre et al. (2018) story segments do not produce any significant change in LM rating behavior. This suggests that the LMs are using different processes than those believed to be used by humans. LMs are able to look past the surface and semantic alterations made by the adversarial attacks.

6 Conclusions

In this study, we investigated the question of whether language models can rate story suspense similarly to humans. Following strict adherence to guidelines on *machine psychology* Hagendorff (2023), we replicated the results of four seminal human suspense judgement experiments. The results are mixed. LMs generally are not capable of providing numerical ratings of suspense that correspond closely to human average numerical ratings. However, it may be the case that one only cares about binary judgments of suspense presence or absence, in which case LMs appear quite competent, at least for the limited selection of stories selected for the original psychology studies. LMs are also much more competent with judgments of rising or falling suspense. This could be useful in applications that help human writers assess how their audiences will perceive their stories O'Neill & Riedl (2009);

Lin & Riedl (2023); Lin et al. (2023). Unfortunately, the points where suspense switches directions from rising to falling, or vice versa, are likely to be the most crucial points in a story, and this is where LMs have the least accuracy in directionality judgments.

We find evidence that LMs do not process suspense the same way that cognitive theories indicate humans process suspense. Adversarial attacks did not have the anticipated effects, and in some cases had the opposite effect. We also find that LMs do not read stories like humans, building up a model of the story across segments as cognitive models of human suspense perception suggest. They appear to focus solely on cues within segments.

7 Ethics

We investigate the extent to which Large Language Models (LMs) align with human perception of suspense in textual narratives. All studies that we compare with our models were obtained in accordance with relevant data usage policies. All authors from previous studies and narratives were cited in our work. No human subjects involved in our survey or experiments outside of computational observations.

A potential harm of our work is the overgeneralization of LM capabilities. Although LMs demonstrate proficiency in text-based analysis of suspense, this does not imply that LMs possess human-like understanding of complex psychological phenomena.

Our studies come from diverse cultural backgrounds—including English, American, Hungarian, Spanish, and German sources; however, our LMs are fed English-translated versions of the original narratives, so subtle linguistic and cultural nuances may be lost. We also want to emphasize that suspense is not universally perceived in the same way. Factors such as age, gender, and socioeconomic status may greatly influence how individuals experience uncertainty and tension in narratives. We wanted to recognize that further research can be done into how these variations can affect how human perception of suspense compares to LMs.

Acknowledgments

The authors collectively would like to thank our anonymous reviewers for their comments and feedback. We appreciate the help provided by Madhura Keshava Ummettuguli with exploring the ideas during the conception of this paper. The costs for some of the model inference in experiments were covered in part using the generous support to Glenn Matlin from TOGETHERAI.

References

Maria Bentz, Maya Cortez Espinoza, Vesela Simeonova, Tilmann Köppe, and Edgar Onea. Measuring suspense in real time: A new experimental methodology. *Sci. Study Lit.*, 12(1): 92–112, 18 April 2024.

E. Branigan. *Narrative Comprehension and Film*. Routledge, New York, 1992.

W F Brewer. The nature narrative suspense problem rereading. In P Vorderer, H J Wulff, and M Friedrichsen (eds.), *Suspense: Conceptualizations, Theoretical Analyses, Empirical Explorations*, pp. 107–127. Lawrence Erlbaum Associates, Mahwah, NJ, 1996.

William F Brewer and Edward H Lichtenstein. Stories are to entertain: A structural-affect theory of stories. *J. Pragmat.*, 6(5-6):473–486, December 1982.

William F Brewer and Keisuke Ohtsuka. Story structure, characterization, just world organization, and reader affect in american and hungarian short stories. *Poetics (Amst.)*, 17(4-5):395–415, October 1988.

N Carroll. Toward theory film suspense. *Persistence Vision*, 1:65–89, 1984.

Yun-Gyung Cheong and Robert Michael Young. Suspenser: A story generation system for suspense. *IEEE Transactions on Computational Intelligence and AI in Games*, 7:39–52, 2015. URL https://api.semanticscholar.org/CorpusID:9179175.

B N Colby, Andrew Ortony, Gerald L Clore, and Allan Collins. The cognitive structure of emotions. *Contemp. Sociol.*, 18(6):957, November 1989.

Pablo Delatorre, Carlos León, Alberto Salguero, Manuel Palomo-Duarte, and Pablo Gervás. Confronting a paradox: A new perspective of the impact of uncertainty in suspense. *Front. Psychol.*, 9:1392, 8 August 2018.

Pablo Delatorre, Carlos León, Alberto G. Salguero, and Alan Tapscott. Predicting the effects of suspenseful outcome for automatic storytelling. *Knowl. Based Syst.*, 209:106450, 2020. URL https://api.semanticscholar.org/CorpusID:224867428.

Pablo Delatorre, Carlos León, and Alberto Salguero Hidalgo. Improving the fitness function of an evolutionary suspense generator through sentiment analysis. *IEEE Access*, 9:39626–39635, 2021. URL https://api.semanticscholar.org/CorpusID:232317642.

Richard Doust and Paul Piwek. A model of suspense for narrative generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

Matthew William Fendt and Robert Michael Young. Leveraging intention revision in narrative planning to create suspenseful stories. *IEEE Transactions on Computational Intelligence and AI in Games*, 9:381–392, 2017. URL https://api.semanticscholar.org/CorpusID:38348125.

G Freytag. *Freytag, G. (1894). Freytag's Technique Drama: Exposition Dramatic Composition Art*. Chicago, IL; Scott Foresman, 1894.

Richard Gerrig. Experiencing narrative worlds: on the psychological activities of reading. *Choice (Middletown)*, 31(07):31–3591–31–3591, 1 March 1994.

Richard J Gerrig and Allan B I Bernardo. Readers as problem-solvers in the experience of suspense. *Poetics (Amst.)*, 22(6):459–472, December 1994.

Sian Gooding, Lucia Lopez-Rivilla, and Edward Grefenstette. Writing as a testbed for open ended agents. *arXiv [cs.CL]*, 25 March 2025.

Arthur C. Graesser, Murray Singer, and Tom Trabasso. Constructing inferences during narrative text comprehension. *Psychological review*, 101 3:371–95, 1994. URL https://api.semanticscholar.org/CorpusID:523868.

J Guidry. The experience of . . . suspense: understanding the construct, its antecedents, and its consequences in consumption and acquisition contexts. 2005.

Thilo Hagendorff. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *ArXiv*, abs/2303.13988, 2023. URL https://api.semanticscholar.org/CorpusID:257757370.

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in german and english poetry. *arXiv [cs.CL]*, 17 March 2020.

Hans Hoeken and Mario van Vliet. Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story. *Poetics (Amst.)*, 27(4): 277–286, May 2000.

Chihmao Hsieh. Suspense: exploring a new lens for outcome uncertainty in entrepreneurship. *Int. J. Entrep. Ventur.*, 13(1):27, 2021.

Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 2014. URL https://api.semanticscholar.org/CorpusID:12233345.

Kai Kaspar, Daniel Zimmermann, and Anne-Kathrin Wilbers. Thrilling news revisited: The role of suspense for the enjoyment of news stories. *Front. Psychol.*, 7:1913, 15 December 2016.

Moritz Lehne and Stefan Koelsch. Toward a general psychological model of tension and suspense. *Front. Psychol.*, 6, 11 February 2015.

Zhi-Wei Li, Neil R Bramley, and Todd M Gureckis. Expectations about future learning influence moment-to-moment feelings of suspense. *Cogn. Emot.*, 35(6):1099–1120, 18 August 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Ré, Diana Acosta-Navas, Drew A Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *arXiv [cs.CL]*, 16 November 2022.

Zhiyu Lin and M. O. Riedl. An ontology of co-creative ai systems. In *Proceedings of the 2024 NeurIPS Workshop on Machine Learning for Creativity and Design*, 2023. URL https://arxiv.org/abs/2310.07472.

Zhiyu Lin, Upol Ehsan, Rohan Agarwal, Samihan Dani, Vidushi Vashishth, and Mark O. Riedl. Beyond prompts: Exploring the design space of mixed-initiative co-creativity systems. In *Proceedings of the 2023 International Conference on Computational Creativity*, 2023. URL https://arxiv.org/abs/2305.07465.

Bing Liu. *Sentiment Analysis and Opinion Mining*. Springer International Publishing, Cham, 2012.

Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. Is machine psychology here? on requirements for using human psychological tests on large language models. *Int Conf Nat Lang Gener*, pp. 230–242, 2024.

Robert Madrigal, Colleen Bee, Johnny Chen, and Monica LaBarge. The effect of suspense on enjoyment following a desirable outcome: The mediating role of relief. *Media Psychol.*, 14 (3):259–288, 31 August 2011.

Simon Malberg, Roman Poletukhin, Carolin M Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs. *arXiv [cs.CL]*, 20 October 2024.

Julie Guidry Moulard, Michael W Kroff, and Judith Anne Garretson Folse. Unraveling consumer suspense: The role of hope, fear, and probability fluctuations. *J. Bus. Res.*, 65(3): 340–346, March 2012.

Brian O'Neill and Mark Riedl. Toward a computational framework of suspense and dramatic arc. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, pp. 246–255, Memphis, Tennessee, October 2011. URL http://www.cc.gatech.edu/~riedl/pubs/acii11.pdf.

Brian O'Neill and Mark Riedl. Dramatis: A computational model of suspense. *Proc. Conf. AAAI Artif. Intell.*, 28(1), 21 June 2014.

Brian O'Neill and Mark O. Riedl. Supporting human creative story authoring with a synthetic audience. In *Proceedings of the 7th Creativity and Cognition Conference*, Berkeley, California, October 2009. URL http://www.cc.gatech.edu/~riedl/pubs/oneill-cc09.pdf.

Samuel J Paech. Creative writing bench. https://eqbench.com/creative_writing.html, 2023. Accessed: 2025-3-27.

L E Palich and D R Bagby. Using cognitive theory explain entrepreneurial risk-taking - challenging conventional wisdom. *Journal Business Venturing*, 10:425–438, 1995.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Erik M Peterson and Arthur A Raney. Reconceptualizing and reexamining suspense as a predictor of mediated sports enjoyment. *J. Broadcast. Electron. Media*, 52(4):544–562, 7 November 2008.

Andrew Piper, Richard Jean So, and David Bamman. Narrative theory for computational narrative understanding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-Tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 298–311, Stroudsburg, PA, USA, November 2021. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. *arXiv [cs.CL]*, 2 March 2024.

Raj Shah, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj, and Sashank Varma. Numeric magnitude comparison effects in large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6147–6161, Toronto, Canada, July 2023. Association for Computational Linguistics.

Ammar Shaikh, Raj Abhijit Dandekar, Sreedath Panat, and Rajat Dandekar. CBEval: A framework for evaluating and interpreting cognitive biases in LLMs. *arXiv [cs.CL]*, 4 December 2024.

Jonathan Shaki, Sarit Kraus, and Michael Wooldridge. Cognitive effects in large language models. *arXiv [cs.AI]*, 28 August 2023.

Max Steg, Karlo H R Slot, and Federico Pianzola. Computational detection of narrativity: A comparison using textual features and reader response. 2022.

A Vaswani, N Shazeer, N Parmar, and others. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 2017.

Peter Vorderer, Hans Jürgen Wulff, and Mike Friedrichsen. *Suspense: Conceptualizations, theoretical analyses, and empirical explorations*. Routledge, London, England, 1996.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv [cs.CL]*, 3 June 2024.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.*, 55(7): 5731–5780, October 2022.

David Wilmot and Frank Keller. Modelling suspense in short stories as uncertainty reduction over neural representation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

Kaige Xie and Mark Riedl. Creating suspenseful stories: Iterative planning with large language models. *arXiv [cs.CL]*, 27 February 2024.

Jichen Zhu, Santiago Ontañón, and Brad Lewter. Representing game characters' inner worlds through narrative perspectives. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, New York, NY, USA, 29 June 2011. ACM.

A Suspense

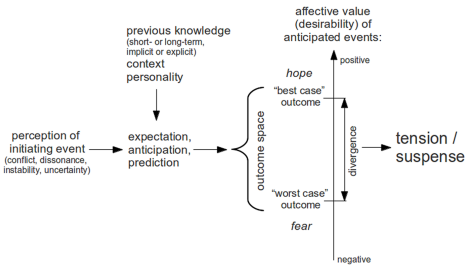


Figure 4: Example of one model of suspense from Lehne & Koelsch (2015) (Figure 2 "Tension Model", pg. 7)

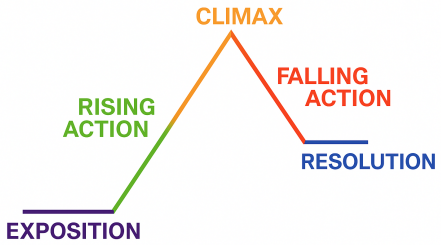


Figure 5: Freytag's Pyramid (Freytag, 1894)

B Adversarial Attacks

Starting from the datasets and human-annotated labels of Gerrig & Bernardo (1994) and Delatorre et al. (2018), we employ 11 different adversarial strategies for text attack. Each model-dataset-attack combination is performed three times. Unless explicitly noted, the suspense scores are averaged over the three generations either as a number or a label. Attacks are performed on a per-passage basis, where a passage is any snippet of text that is followed by a request for a suspense rating. For the Gerrig & Bernardo (1994) dataset, the entire passage and question are shown in a single sequence. Delatorre et al. (2018), on the other hand, is comprised of 12 different segments, which are each attacked individually. The following excerpts are examples of each text attack. As a consistent example, we use the first paragraph from the stories from Delatorre et al. (2018) showing the resulting text after each attack method is applied.

Control

```
"1 What you are about to read is the story of a real event that took place at the
UCSF Benioff Children's Hospital, in San Francisco, California, on 24 February
2008. On that day, since 8 a.m., an eight year old boy called Robert Bent and the
entire medical team treating him were all ready for his imminent liver transplant.
Just the day before a suitable donor had been found, and they were now awaiting
the arrival of the organ. However, they were not sure if Robert would survive
the wait as his situation was critical. This is the story of what happened."
```

Word Swapping

```
"1 Francisco, that What event in 24 story took read of February at San about is
place Benioff you the California, to Children's UCSF Hospital, are a on real the
2008. On and old since 8 year a.m., team Robert for an all eight Bent day, the
entire that called transplant. medical boy him ready imminent his treating were
liver they donor of found, before the the the had Just day awaiting and a organ.
arrival suitable were now been they were Robert his survive sure situation was
the critical. would as However, not wait if happened. story is This what the of"
```

Antonym Substitution

```
"1 What you differ about to read differ the story of a nominal event that give
place at the UCSF Benioff Children's Hospital, in San Francisco, California, on
24 February 2008. On that day, since 8 a. m. , an eight year young boy called
Robert Bent and the entire surgical team treating him differ all unready for his
```

imminent liver transplant. Just the day before a suitable donor lack differ lose, and they differ now awaiting the arrival of the organ. However, they differ not uncertain if Robert would succumb the wait as his situation differ noncritical. This differ the story of what dematerialize."

Synonym Substitution

"1 What you live astir to study be the account of a substantial event that took plaza at the UCSF Benioff Nestling 's Infirmary, in San Francisco, Calif., on 24 Feb 2008. On that day, since 8 a. m. , an ashcan school class old boy foretell Robert Bent and the entire medical squad treating him were all quick for his imminent liver graft. Just the day before a suited donor had equal found, and they be at present awaiting the arrival of the electronic organ. However, they were non trusted if Henry m. robert would live on the wait as his berth comprise decisive. This is the write up of what happen."

Caesar Ciphering

"1 Zkdw brx duh derxw wr uhdg lv wkh vwrub ri d uhdo hyhqw wkdw wrrn sodfh dw wkh XFVI Ehqlrii Fkloguhq'v Krvslwdo, lq Vdq Iudqflvfr, Fdoliruqld, rq 57 Iheuxdub 5331. Rq wkdw gdb, vlqfh 1 d.p., dq hljkw bhdu rog erb fdoohg Urehuw Ehqw dqg wkh hqwluh phglfdo whdp wuhdwlqj klp zhuh doo uhdgb iru klv lpplqhqw olyhu wudqvsodqw. Mxvw wkh gdb ehiruh d vxlwdeoh grqru kdg ehhq irxqg, dqg wkhb zhuh qrz dzdlwlqj wkh duulydo ri wkh rujdq. Krzhyhu, wkhb zhuh qrw vxuh li Urehuw zrxog vxuylyh wkh zdlw dv klv vlwxdwlrq zdv fulwlfdo. Wklv lv wkh vwrub ri zkdw kdsshqhg. "

Homoglyph Substitution

"1 What **y**ou are abo**u**t **t**o **re**ad is the story of a real event tha**t** took p**l**a**c**e **a**t **the** UCSF Be**n**ioff Children$\check{2}$019**s** Hospital, in San Franc**i**s**c**o, California, on 24 Febru**a**ry 200**8**. On th**a**t da**y**, s**in**ce **8** a.m., an eight ye**a**r old boy called Rob**e**r**t** B**e**nt and the entire medical team treati**n**g **h**im were a**l**l ready for his immi**n**ent **l**iver t**r**ansplan**t**. Just the da**y** before a suitable donor had b**ee**n found, and the**y** were now awa**i**ting th**e** arriva**l** of the organ. Howe**ver**, they were not sure if Robert w**o**u**l**d **s**urvive the wait as his situatio**n** w**as** criti**c**a**l**. T**h**is is the story of what happened."

Distraction Phrase Insertion

"1 What you are about to read is the story of a real event that took place at the UCSF Benioff Children's Hospital, in San Francisco, California, on 24 February 2008. On that day, since 8 a. m. , an eight year old boy called Robert Bent and the entire medical team treating him were all ready for his imminent liver transplant. He looked for his hidden watch. He couldn't find it. Just the day before a suitable donor had been found, and they were now awaiting the arrival of the organ. However, they were not sure if Robert would survive the wait as his situation was critical. This is the story of what happened. ",

Sentence Shuffling

"1 The analysis showed that it had withstood the impact and it was possible to use the organ for the transplant. When they opened the case, they discovered that the interior bag had ruptured. The two men transporting the liver left the roof via the doorway to the service stairwell, which they decided to walk down. The doctors arrived promptly. However, they were not sure if Robert would survive the wait as his situation was critical."

Embedding-Based Word Substitution

"1 Whereof you constituted roughly dans read is du histories des other reales phenomena that picked site into the UCSF Benioff Children'r Clinic, in Santo Stefania, California, orn 24 February 2008. Orn somebody daytime, because 8 a.m., an seis annual antigua guy titled Robert Bent and per whole medical squad treating he were all ready in his imminent renal transplanted. Virtuous de day beforehand latest suitable donor has played uncovered, ja would were presently awaiting by arriving de nova agency. However, would was nope sure if Richards would live the awaits como her circumstances es crucial. This is de story of what transpired."

Character Name Changes

"1 What you are about to read is the story of a real event that took place at the UCSF Benioff Children's Hospital, in San Francisco, California, on 24 February 2008. On that day, since 8 a.m., an eight year old boy called Alex and the entire medical team treating him were all ready for his imminent liver transplant. Just the day before a suitable donor had been found, and they were now awaiting the arrival of the organ. However, they were not sure if Alex would survive the wait as his situation was critical. This is the story of what happened. "

Context Removal

"1 What you are about to read is the story of a real event that took place at the UCSF Benioff Children's Hospital, in San Francisco, California, on 24 February 2008. On that day, since 8 a.m., an eight year old boy called Robert Bent and the entire medical team treating him were all ready for his imminent liver transplant. Just the day before a suitable donor had been found, and they were now awaiting the arrival of the organ. This is the story of what happened. "

Typo Insertion

"1 What you are AF8uF to read is the story of a real event rtWY HK9M 0kZxe at the UCSF Benioff Children's Hospital, in San Francisco, XakkT8rBlW, on 24 e4fFHzEy 2008. On 6uqG day, since 8 a. m. , an eight year old boy called Robert Bent and the entire medical team treating him were all ready for his imminent liver transplant. Just the day before a suitable donor had been found, and they A2tR now awaiting the arrival of the organ. However, they were not dHfW if Robert would survive the wait as his situation was critical. GNJa is the story of what happened. "

C Results In Detail

This Appendix gives full-size charts and additional tables for the main results in Section §4: RESULTS.

	Rising Action						Falling Action					
	Delatorre			Lehne			Delatorre			Lehne		
Model	**Acc.** ↑	**F1** ↑	**MSE** ↓	**Acc.** ↑	**F1** ↑	**MSE** ↓	**Acc.** ↑	**F1** ↑	**MSE** ↓	**Acc.** ↑	**F1** ↑	**MSE** ↓
Qwen2 72B	0.94 ± 0.18	0.96 ± 0.12	2.31 ± 0.94	0.62 ± 0.10	0.71 ± 0.09	2.66 ± 0.26	0.73 ± 0.00	0.84 ± 0.00	1.66 ± 0.00	0.84 ± 0.00	0.90 ± 0.00	1.98 ± 0.00
DeepSeek V3	0.94 ± 0.18	0.96 ± 0.12	1.83 ± 0.81	0.63 ± 0.13	0.74 ± 0.09	2.29 ± 0.18	0.86 ± 0.00	0.91 ± 0.00	1.66 ± 0.00	0.86 ± 0.00	0.91 ± 0.00	1.77 ± 0.00
Gemma 2 27b	1.00 ± 0.00	1.00 ± 0.00	1.69 ± 0.68	0.74 ± 0.06	0.81 ± 0.04	2.02 ± 0.09	0.86 ± 0.00	0.92 ± 0.00	1.88 ± 0.00	0.86 ± 0.00	0.92 ± 0.00	2.17 ± 0.00
Gemma 2 9b	1.00 ± 0.00	1.00 ± 0.00	1.96 ± 0.70	0.62 ± 0.08	0.76 ± 0.05	2.44 ± 0.32	0.77 ± 0.00	0.86 ± 0.00	1.83 ± 0.00	0.86 ± 0.00	0.91 ± 0.00	2.27 ± 0.00
Llama 2 7b	1.00 ± 0.00	1.00 ± 0.00	2.77 ± 0.73	0.63 ± 0.04	0.77 ± 0.03	3.31 ± 0.14	0.86 ± 0.00	0.93 ± 0.00	3.52 ± 0.00	0.81 ± 0.00	0.90 ± 0.00	3.20 ± 0.00
Llama 3 70b	0.81 ± 0.37	0.83 ± 0.36	3.20 ± 0.98	0.69 ± 0.09	0.76 ± 0.09	3.13 ± 0.30	0.73 ± 0.00	0.81 ± 0.00	2.56 ± 0.00	0.79 ± 0.00	0.86 ± 0.00	2.53 ± 0.00
Llama 3 8b	0.69 ± 0.37	0.75 ± 0.35	3.14 ± 0.40	0.64 ± 0.08	0.75 ± 0.07	2.82 ± 0.24	0.59 ± 0.00	0.71 ± 0.00	3.01 ± 0.00	0.79 ± 0.00	0.86 ± 0.00	2.48 ± 0.00
WizardLM 2	1.00 ± 0.00	1.00 ± 0.00	2.20 ± 1.16	0.64 ± 0.10	0.75 ± 0.10	2.51 ± 0.60	0.77 ± 0.00	0.86 ± 0.00	1.90 ± 0.00	0.86 ± 0.00	0.91 ± 0.00	1.67 ± 0.00
Mistral 7B	0.94 ± 0.18	0.96 ± 0.12	2.53 ± 0.86	0.63 ± 0.08	0.73 ± 0.07	2.60 ± 0.47	0.86 ± 0.00	0.92 ± 0.00	3.21 ± 0.00	0.84 ± 0.00	0.79 ± 0.00	2.89 ± 0.00
Mixtral 8x7B	1.00 ± 0.00	1.00 ± 0.00	2.11 ± 0.92	0.63 ± 0.08	0.75 ± 0.05	3.01 ± 0.41	0.68 ± 0.00	0.77 ± 0.00	2.33 ± 0.00	0.72 ± 0.00	0.79 ± 0.00	1.89 ± 0.00
Average	0.93 ± 0.13	0.95 ± 0.11	2.37 ± 0.82	0.65 ± 0.08	0.75 ± 0.07	2.68 ± 0.30	0.77 ± 0.00	0.85 ± 0.00	2.36 ± 0.00	0.82 ± 0.00	0.89 ± 0.00	2.29 ± 0.00

Table 3: Brewer, and Lehne results on periods of rising or falling action when binarized the numerical scale.

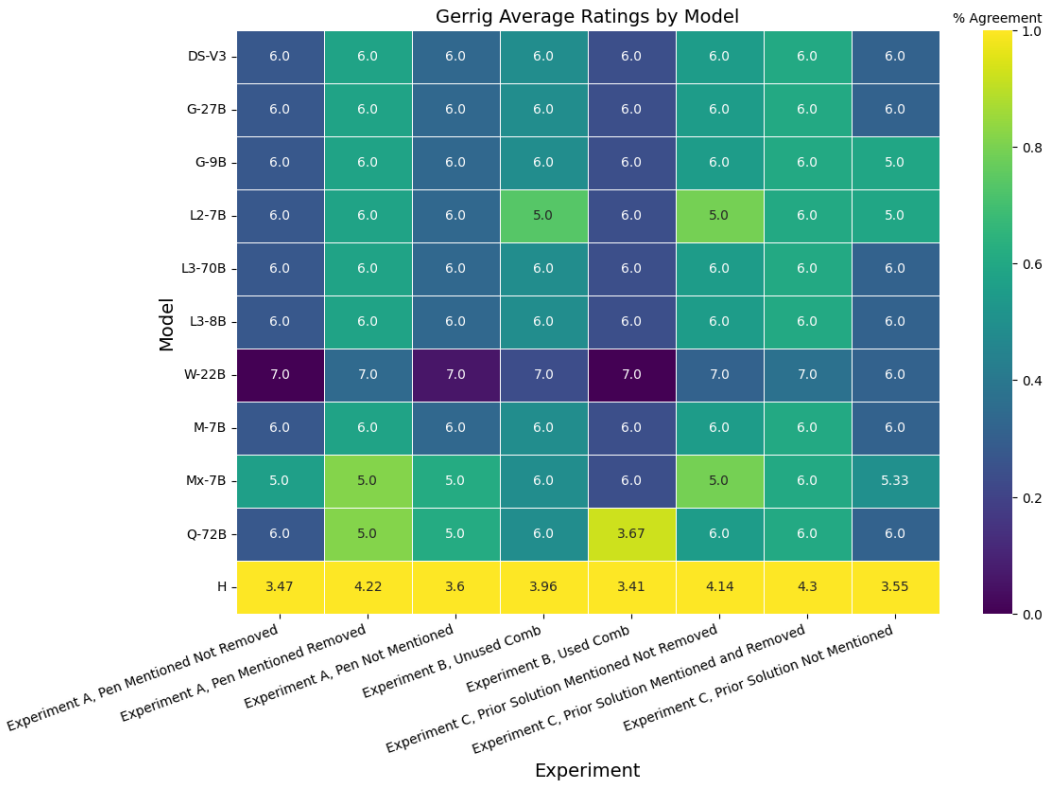
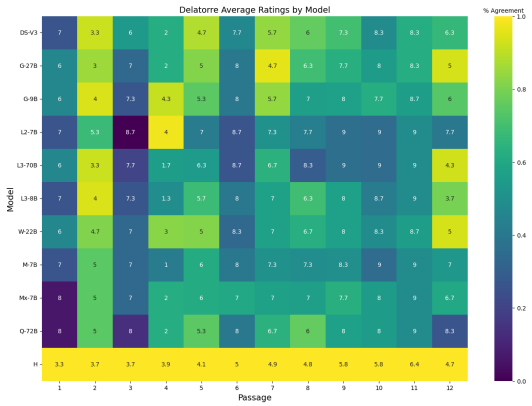
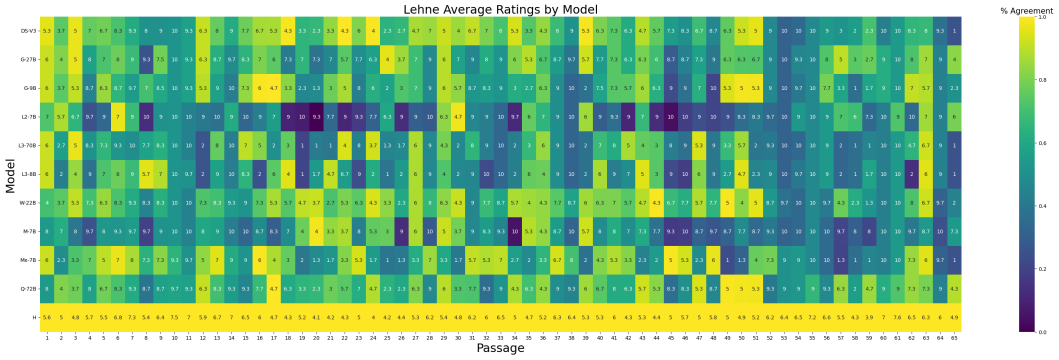


Figure 6: Gerrig results heatmap. Rows are models. Columns are different treatments of the story.

(a) Brewer results (Birthday story).

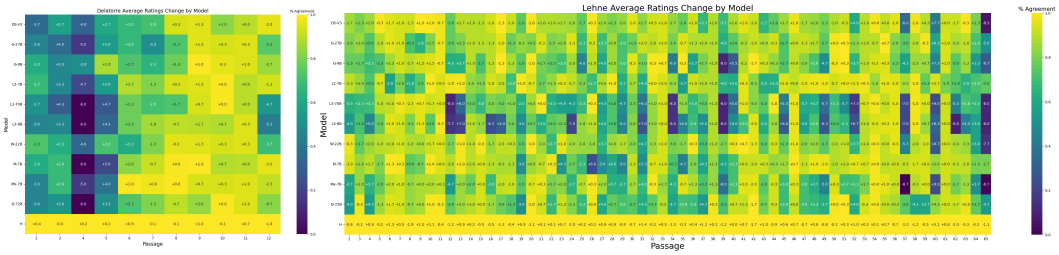


(b) Delatorre results.



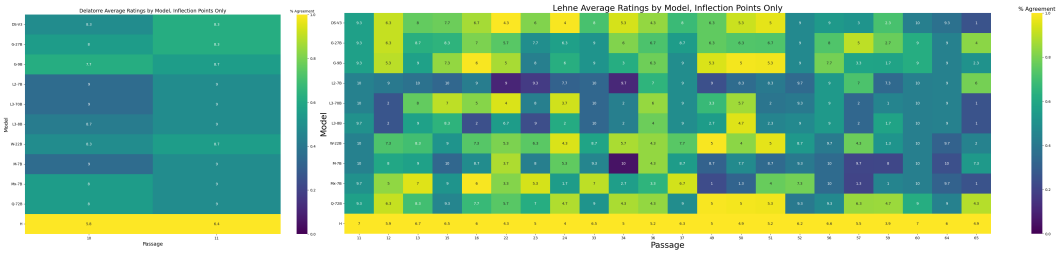
(c) Lehne results.

Figure 7: Results for rating agreement. Rows are models, with bottom row the average ratings across all LMs. Columns are story segments.



(a) Delatorre directionality.

(b) Lehne directionality.



(c) Delatorre inflection pts.

(d) Lehne inflection points.

Figure 8: Directionality results for Delatorre and Lehne—whether the model predicts rising or falling suspense across segments in accordance with human ratings rising or falling. Bottom graphs isolate segements where rising suspense changes to falling, and vice versa.

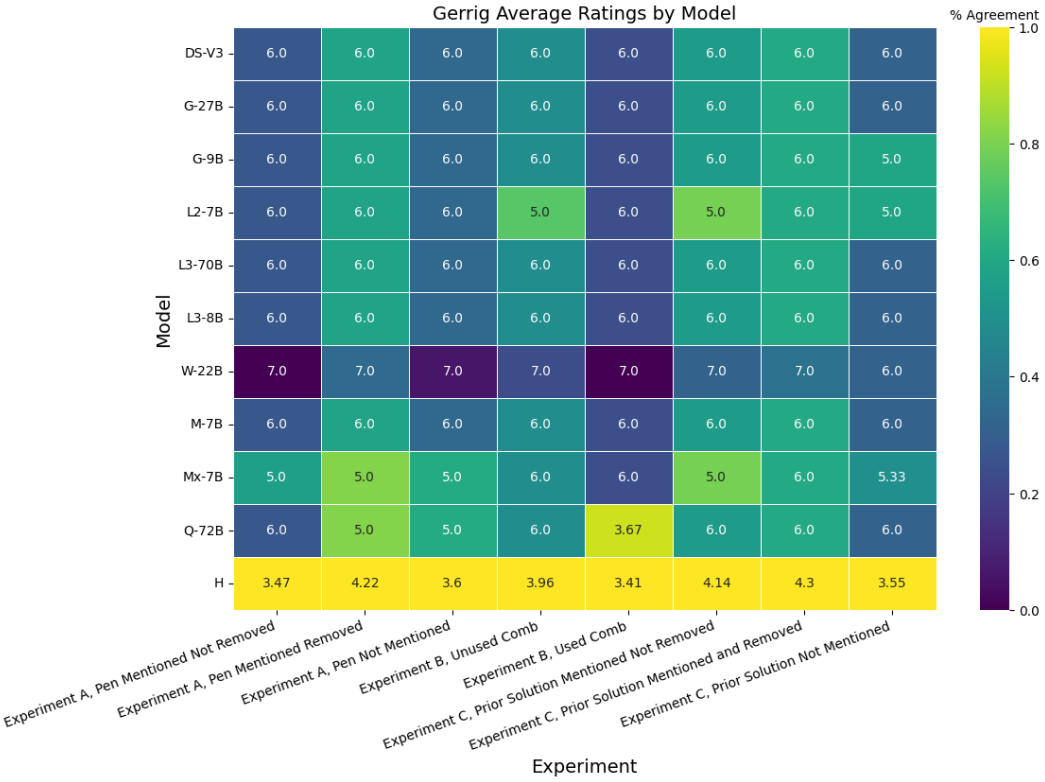


Figure 9: Gerrig average rating agreements.

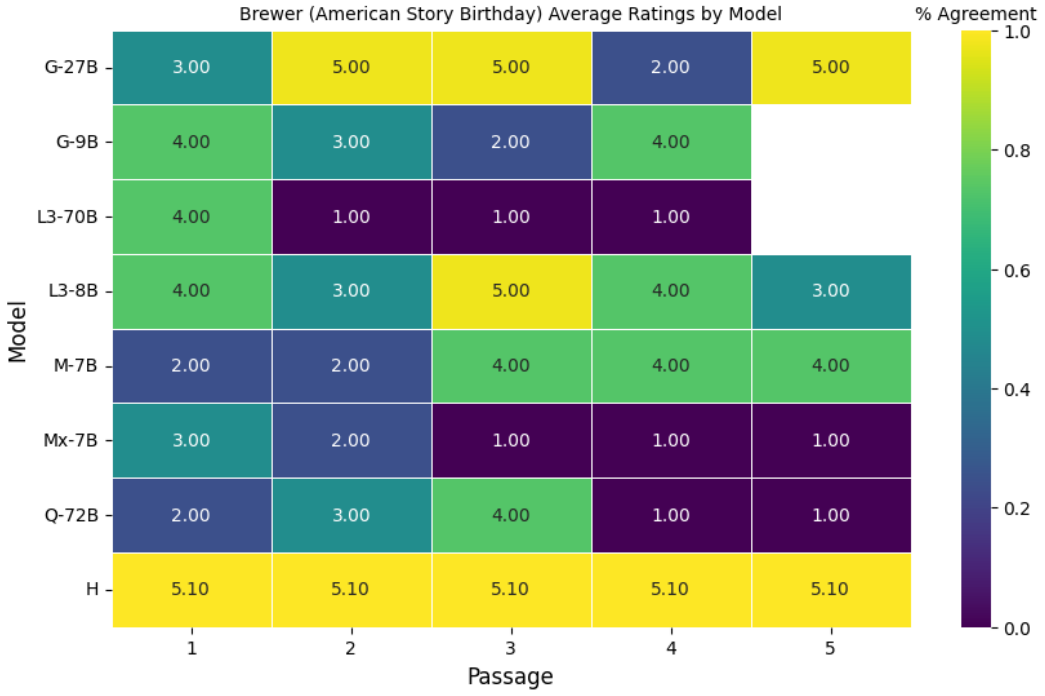


Figure 10: Brewer average rating agreements for the Birthday story.

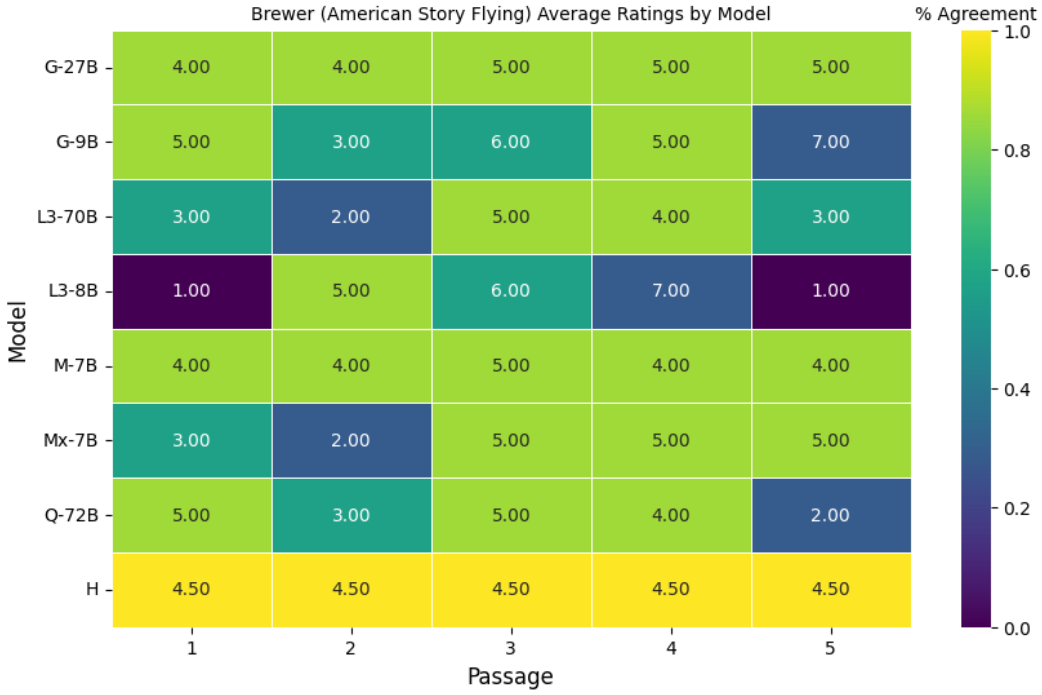


Figure 11: Brewer average rating agreements for the Flying story.

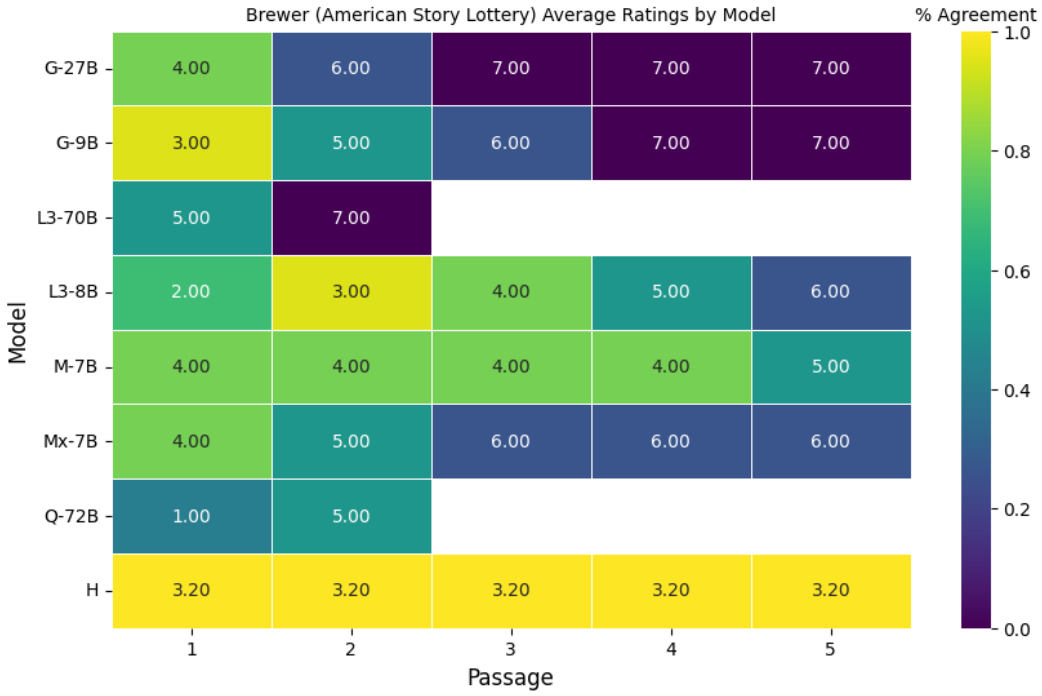


Figure 12: Brewer average rating agreements for the Lottery story.

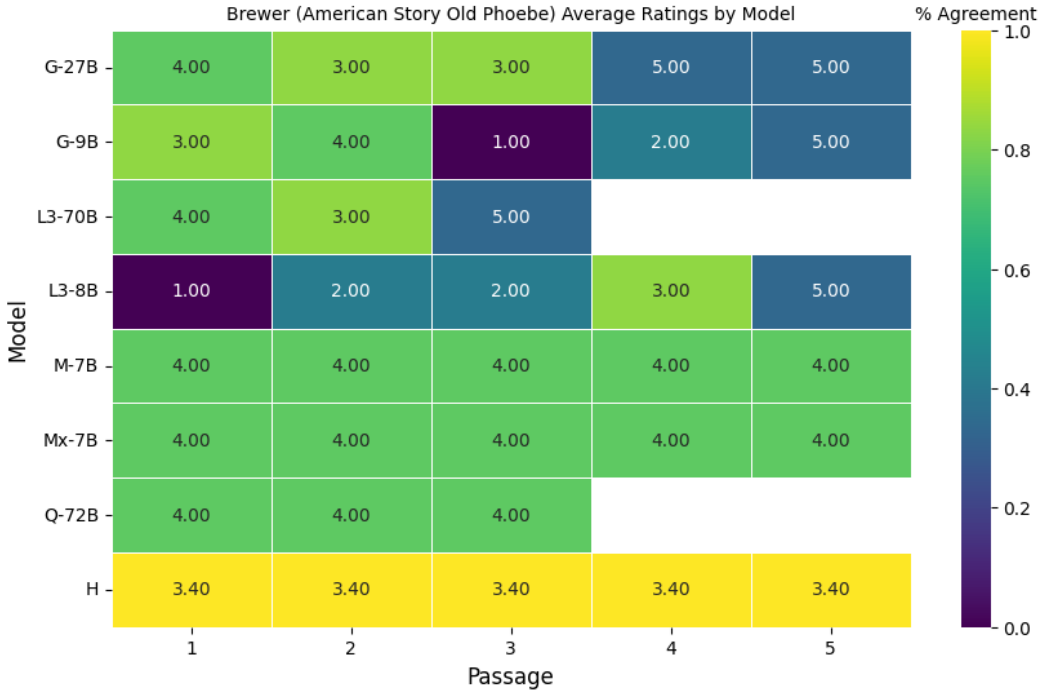


Figure 13: Brewer average rating agreements for the Old Phoebe story.

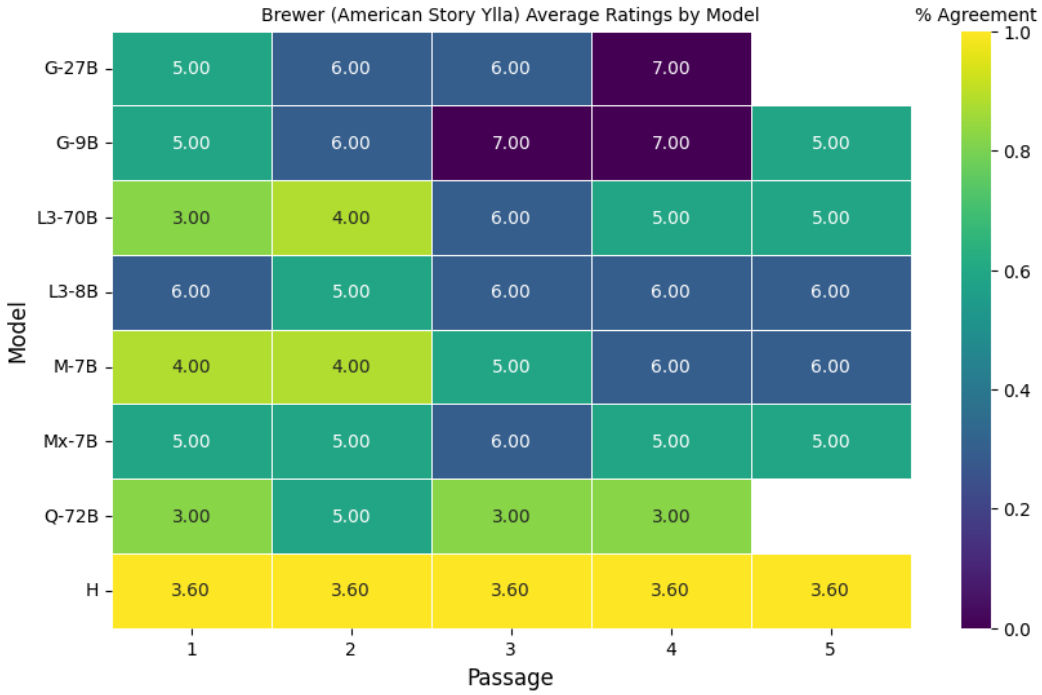


Figure 14: Brewer average rating agreements for the Ylla story.

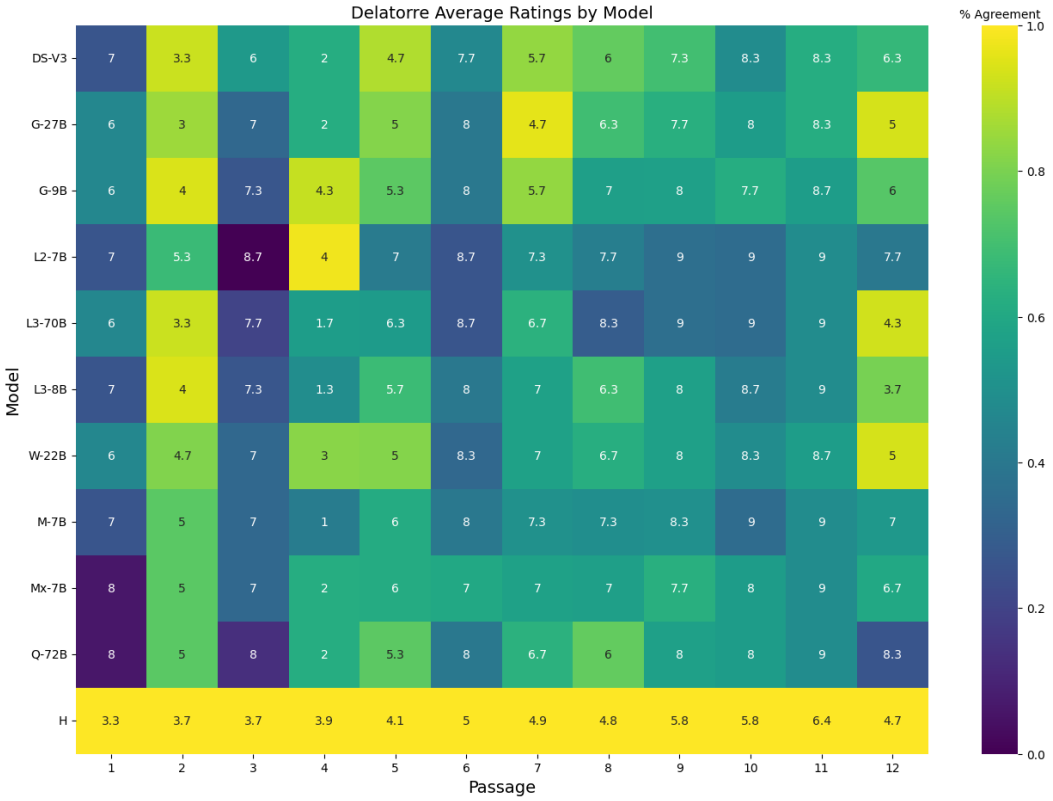


Figure 15: Delatorre average rating agreements.

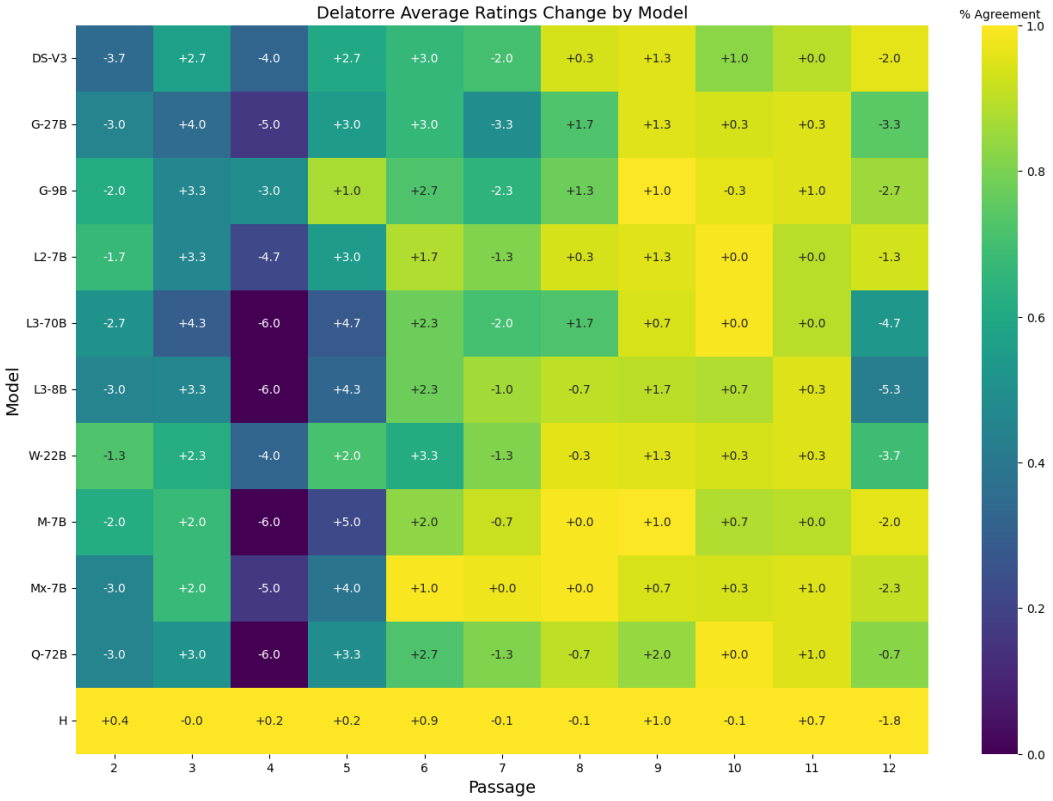


Figure 16: Delatorre average rating change agreements.

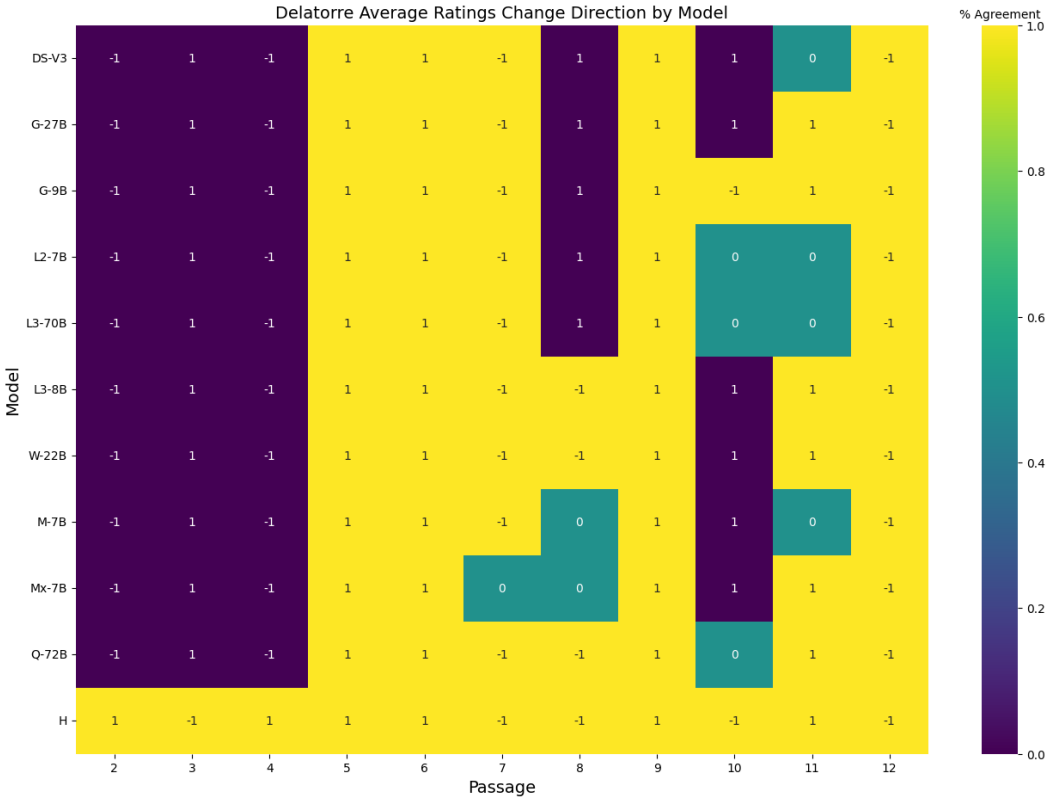


Figure 17: Delatorre average rating change direction agreements.

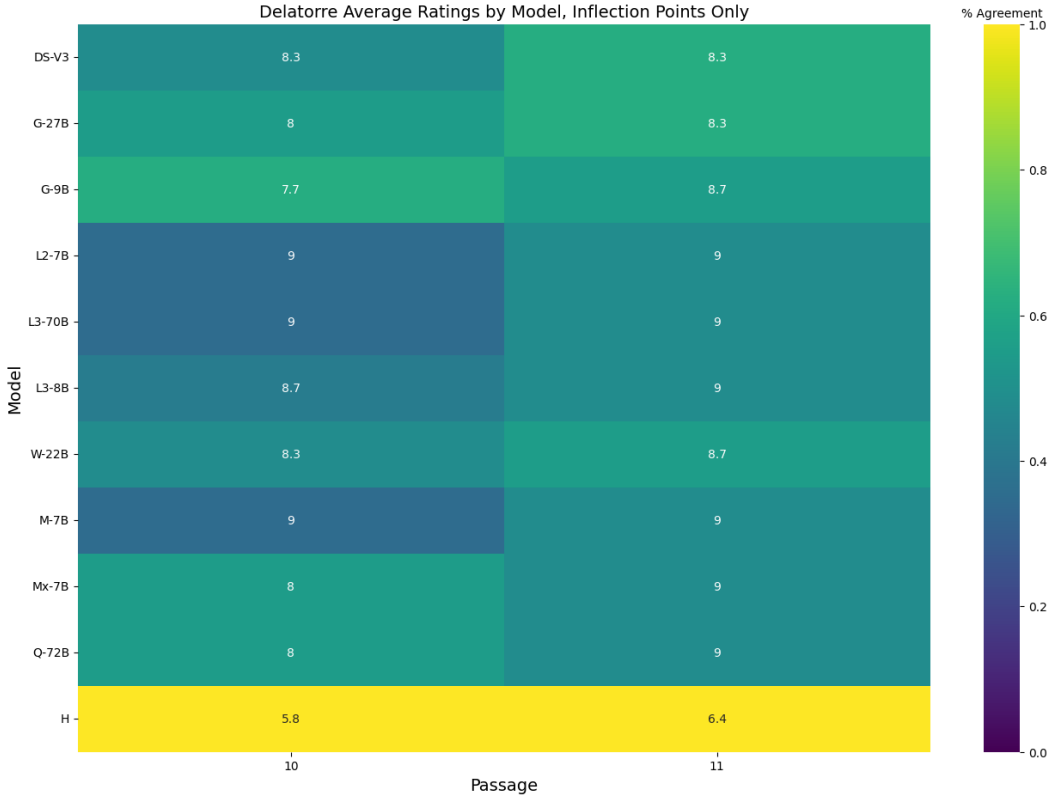


Figure 18: Delatorre average rating agreements, only passages where action changes to rising or falling.

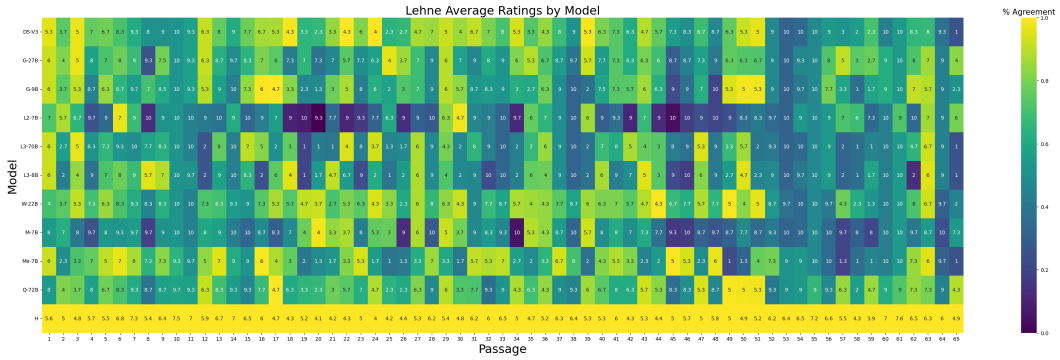


Figure 19: Lehne average rating agreements.

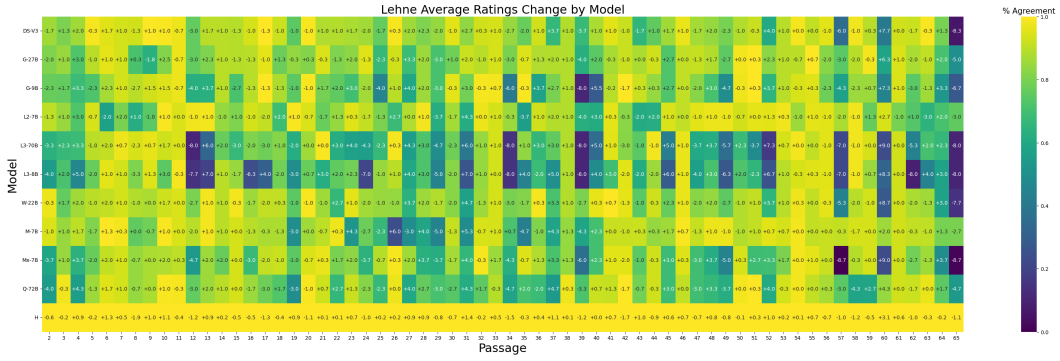


Figure 20: Lehne average rating change agreements.

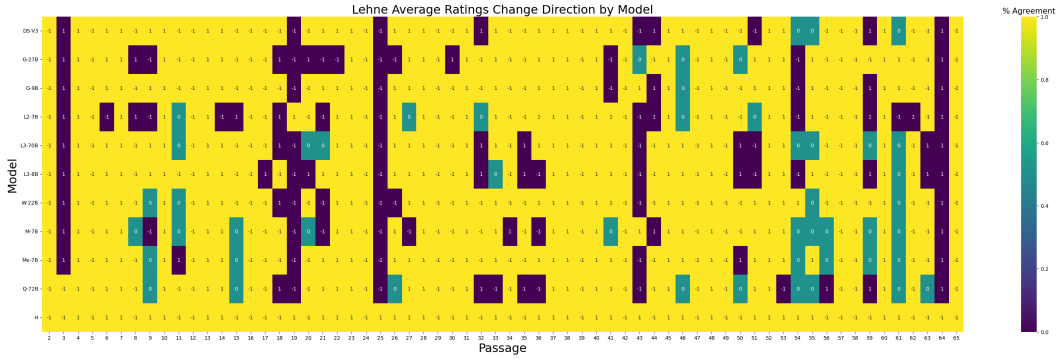


Figure 21: Lehne average rating change direction agreements.

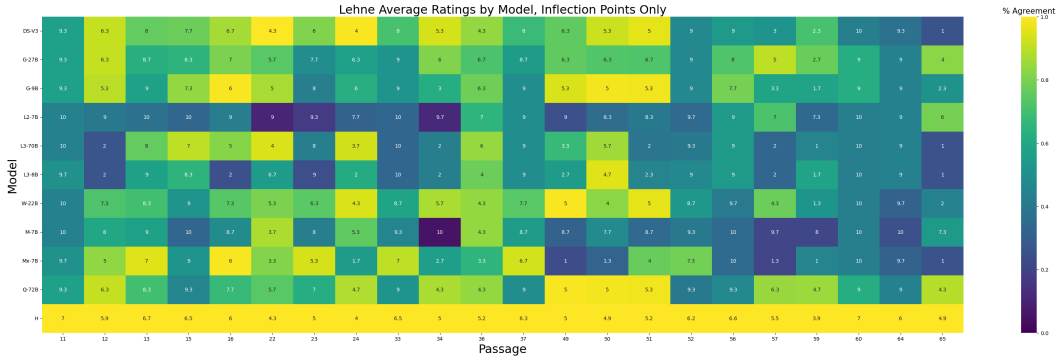


Figure 22: Lehne average rating agreements, only passages where action changes to rising or falling.