Tracking Evolving Relationship Between Characters in Books in the Era of Large Language Models

Abhilasha Sancheti **Rachel Rudinger**
University of Maryland, College Park
{sancheti, rudinger}@umd.edu

Abstract

This work aims to assess the zero-shot social reasoning capabilities of LLMs by proposing various strategies based on the granularity of information used to track the fine-grained evolution in the relationship between characters in a book. Without gold annotations, we thoroughly analyze the agreements between predictions from multiple LLMs and manually examine their consensus at a local and global level via the task of trope prediction. Our findings reveal low-to-moderate agreement among LLMs and humans, reflecting the complexity of the task. Analysis shows that LLMs are sensitive to subtle contextual changes and often rely on surface-level cues. Humans, too, may interpret relationships differently, leading to disagreements in annotations.

1 Introduction

Plots and characters are the two key components of a narrative (among others) that contribute to a good piece of fiction (Kennedy and Gioia, 1983; McKee, 1997; Card, 1999). Character comprehension is key to understanding narratives in literary, and psychological research (Bower and Morrow, 1990; Paris and Paris, 2003; Currie, 2009; Kennedy et al., 2013). Particularly, characters and their relationships are one of the basic building blocks of narratives that make them engaging and interesting. Such relationships develop chapter-by-chapter in response to various events as the story progresses. For instance, Figure 1 depicts how Jana and Anil's relationship in *Jana Goes Wild* by Farah Heron, evolves from intense love to a painful breakup, followed by a separation and re-evaluation of their relationship to fall in love again.

Humans build mental models for characters and keep updating them as they read a narrative to explain such developing relationships, character's identity, their emotional status (Gernsbacher et al., 1998), and future behaviors (Fiske et al., 1979; Mead, 1990). However, a lot of manual hours are spent to obtain such insights. Having an automated system that can predict such insights has many practical applications that include book recommendation systems based on similar or diverse relation-archetype narratives, question-answering systems that can aid readers in recalling the relation-archetypes until a point in the book, and systems to predict character's personality or next action based on nature of the relationship.

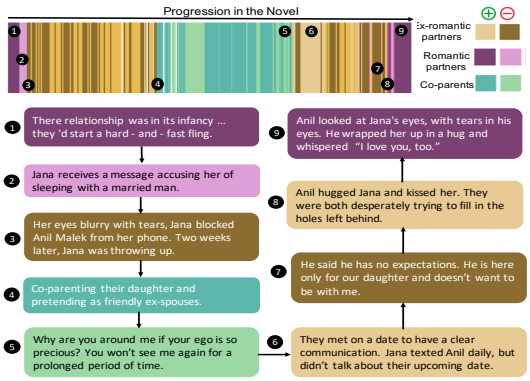


Figure 1: Sample trajectory of evolution in relationship between *Jana* and *Anil* in the book *Jana goes Wild*. Jana and Anil start as romantic partners followed by a tumultuous break-up but years later, co-parenting responsibilities of their daughter force them to confront lingering feelings, reevaluate their past, and rediscover love through shared growth and proximity. ⊕ (⊖) denote a positive (negative) evolution in the relationship.

While there exist works that predict static relationships from movie dialogues (Jia et al., 2021), TV series (Tigunova et al., 2021; Jurgens et al., 2023), or book summaries (Srivastava et al., 2016) and dynamic relationships from book summaries (Chaturvedi et al., 2016, 2017) or a sequence of passages from books (Iyyer et al., 2016), efforts are limited due to the modeling capacity, and unavailability of annotated datasets. With the advent of LLMs, known for their zero/few-shot reasoning capabilities (Brown et al., 2020; Touvron

Proceedings of the The 7th Workshop on Narrative Understanding, pages 64–82

et al., 2023; Jiang et al., 2023) and increased context window size (Team et al., 2023; Dubey et al., 2024), in this work, we ask how can we: (1) characterize evolution in the relationship between characters in book-length text? (2) use LLM's zero-shot reasoning capabilities (without gold labels) to track evolution in the relationship between characters?

We first characterize evolution in the relationship in terms of predefined relationship types and different ways (such as positive, negative, or stable) in which a relationship can evolve (§2). Then, we formally define the task of tracking evolution in the relationship (see Figure 1) between two characters (§3), and propose several strategies based on the granularity of information provided to LLMs to perform the task (§4). To address the issue of the unavailability of gold labels, we evaluate the predictions from the proposed strategies by analyzing the agreement between predictions from different families of LLMs (§7). Low-to-moderate agreement ($\alpha = 0.1 - 0.6$) between predictions from multiple LLMs suggests that the task is difficult even for LLMs with increased context window. To provide an upper bound on the performance achievable from the proposed strategies for this task, we manually examine the consensus predictions at both local and global-level (§8). Low-to-moderate agreement between humans reinforces the difficulty of the task. Finally, we present a quantitative (§9.1) and qualitative analysis (§9.2) of the predictions from LLMs and disagreement between humans to shed light on the challenging nature of this task.

2 Characterizing Evolution in Interpersonal Relationships

Prior works that model the evolution in the relationship between characters use an ontology of relationships that is either coarse-grained (cooperative vs non-cooperative) (Srivastava et al., 2016; Chaturvedi et al., 2016) or unsupervised (Chaturvedi et al., 2017; Iyyer et al., 2016) (such as topics from a topic model). However, relationships can be of various types such as familial (*e.g.*, parent and siblings), social (*e.g.*, friends and acquaintance), romantic (*e.g.*, married and engaged), and professional (*e.g.*, boss and colleague) (Rashid and Blanco, 2018; Tigunova et al., 2021; Jurgens et al., 2023). Following Jurgens et al. (2023), we use a subset of relationship types (see Table 1), that are observed in the most frequently used *tropes*[1] (*e.g.*, enemies-to-lovers, and friends-to-lovers) in romance novels where relationships evolve with time (Lissauer, 2014). Furthermore, relationships are defined by multiple interrelated *interactions* (Blumstein and Kollock, 1988), and the fine-grained characteristics of interactions are not necessarily the same as those of a relationship (*e.g.*, two friends can have a heated argument during an interaction but that does not affect the long-term friendship). Such fine-grained characteristics of interactions and relationships are called *dimensions* in social science (Wish et al., 1976; Deri et al., 2018; Qamar et al., 2021). Inspired by this, we define the interactions between characters using a set of dimensions (such as similarity, trust, romance, social support, identity, respect, knowledge exchange, power, fun, and conflict) proposed by Deri et al. (2018). We believe that change in the intensity of such dimensions determines the *fine-grained* evolution in relationships which can of three types: **positive**, **negative**, and **stable**. A positive evolution signifies deepening connection, increasing trust, support or respect, spending more time together, and sharing similar goals. Any tension in a relationship due to conflicts, arguments, distrust, disrespect, lack of support, or misunderstandings denotes negative evolution. A stable relationship neither evolves positively nor negatively.

Category	Relationship Types
Romantic	Engaged, Married, Romantic interest, Dating, One-sided romantic interest, Separated, Ex-romantic interest, Ex-engaged
Social	Stranger, Acquaintance, Friend, Best friend
Anti-Social	Competitor or Enemy

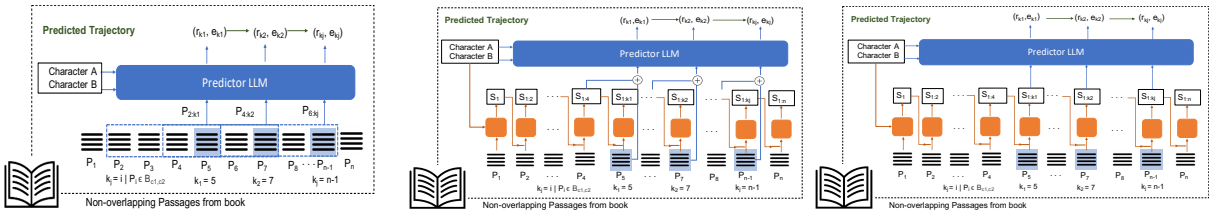
Table 1: The **ontology of relationships** used following prior work (Jurgens et al., 2023).

3 Task of Tracking Evolution in Relationship

We consider tracking evolution in the relationship between characters as a classification task formally defined as follows. Consider a book $B = \{P_1, P_2, \ldots P_n\}$ consisting of n chronologically ordered[2] (in book's passages) non-overlapping passages of a fixed length, c_1 and

[1]Trope refers to a recurring plot device, character archetype, or theme that is commonly used in books.

[2]Please note that we assume a temporally linear plot structure, and leave the modeling of nonlinear timelines (or other complex structures, like worlds within worlds, etc.) for future work (Pustejovsky et al., 2003; Vashishtha et al., 2019).



(a) Complete Passages (b) Summary with Passage (c) Complete Summary

Figure 2: Proposed strategies varying in the granularity of passages provided to an LLM for predicting the evolution in the relationship between given characters. ≡: Passage where both characters are mentioned ■: Summarizer LLM.

c_2 as the two characters, and $B_{c_1,c_2} = \{P_i \in B \mid \text{both } c_1 \text{ and } c_2 \text{ are mentioned in } P_i\}$. Note that B_{c_1,c_2} is a non-contiguous sub-sequence of $P_1, \ldots P_n$. The task is to predict a tuple (r_i, e_i) where $r_i \in R$ and $e_i \in E$, respectively, denote the type of relationship and evolution (from a predefined set as described in §2) between the two characters by the end of the passage $P_i \in B_{c_1,c_2}$ given the passages $P_{1:i}$. We define evolution in the relationship between c_1 and c_2 in a book B as a trajectory $T_{c_1,c_2} = \{(r_1, e_1), (r_2, e_2), \ldots, (r_j, e_j)\}$ of relationship[3] and evolution types at each passage P_{k_j} where $k_j = \{i \mid P_i \in B_{c_1,c_2}\}$.

4 Proposed Strategies for Tracking Evolution in Relationship

Automatic tracking of fine-grained evolution in the relationship between characters in a book-length context poses two main challenges: (1) handling long context, and (2) lack of annotated datasets. To address these challenges, we aim to assess the zero-shot social reasoning capabilities of recent large language models (Jiang et al., 2023; Team et al., 2024; Dubey et al., 2024) with increased context window size by proposing various strategies (Figure 2) based on the granularity of information (*i.e.*, passages $P_{1:i}$) provided to an LLM to predict a relationship and evolution type by the end of $P_i \in B_{c_1,c_2}$ (as defined in §3).

Complete Passages. This strategy uses the large context window of LLMs to provide passages until $P_i \in B_{c_1,c_2}$ (that can fit in the window) as-is in its highest granularity to predict the status of relationship and evolution type until P_i.

Summary with Passage. As books can be arbitrarily long, $P_{1:i} \in B_{c_1,c_2}$ may not always fit in the context window of LLMs. Further, a reader may know the relationship either because the text in passage P_i reveals information about it directly; or because they recall it from prior passages, and no new information is introduced to change or contradict it; or relevant information is introduced in the passage P_i that is best understood in the context of information presented in prior passages. Thus, we hypothesize that providing a "memory" of prior passages is sufficient for relationship type prediction. However, evolution type changes are defined for each interaction between the two characters making it a more granular and local characteristic of relationships. Hence, instead of providing $P_{1:i}$ as-is, this strategy uses a summary of the type and nature of evolution in the relationship between two characters for passages $P_{1:i-1}$[4] (see §4.1) along with the passage P_i to predict the status of the relationship and evolution type by the end of P_i.

Complete Summary. To study if this task can be performed solely with a summary, in this strategy, we provide the complete context until passage P_i as a summary of the type and nature of evolution in the relationship between the two characters.

4.1 Iterative Summary Generation

To obtain the required summary in the above strategies, following Chang et al. (2023b) and Stiennon et al. (2020), we prompt an LLM (in a zero-shot setting) to iteratively generate a summary and update it with every new passage. Formally, $\mathcal{S}(P_{1:i}) = \mathcal{S}(\mathcal{S}(P_{1:i-1}), P_i)$ where, $\mathcal{S}$ is the summarizer LLM, $\mathcal{S}(P_{1:i-1})$ is the previous summary until passage P_{i-1} and $\mathcal{S}(P_{1:i})$ denotes the updated summary until passage P_i. As summaries may exceed a word limit, following Chang et al. (2023b), we repeatedly prompt LLM to compress (Prompt A.3) the summary until it is within the

[3]We consider the presence of one relationship type at one point in this work however, we acknowledge that multiple relationship types may relate two characters at the same time.

[4]Note that P_{i-1} denote the previous passage as per the chronology in B and not B_{c_1,c_2}.

word limit. Generating a summary iteratively allows for the use of LLMs with smaller context windows, making the process faster, less expensive, and more efficient in terms of inference time and number of generated tokens. We provide details on the prompts in §A.2 in appendix.

4.2 Relationship and Evolution Prediction

Given the input for each of the described strategies, we iteratively prompt an LLM to first determine the relationship type and then the evolution type for the chosen relationship in a zero-shot setting. In addition to the predefined set of relationship and evolution types in §2, we also allow LLMs to predict *cannot be determined* for both the tasks and *others* for the relationship type to cover instances when a relationship type may be determined but is not provided in the predefined set. We provide the prompts used for each strategy (§A.2) and other implementation details (§A.1) in appendix.

5 Experimental Setup

We provide details on the source of dataset, preprocessing steps, predictor and summarizer LLMs.

5.1 Dataset Source

While many books are available on resources like Project Gutenberg[5] (Stroube, 2003), LLMs have memorized them along with their summaries available on online sources as study guides[6] (Chang et al., 2023a). Using these books might result in data contamination therefore, we use 11 books (published in 2023) collected by Chang et al. (2023b) that are less likely to be memorized by LLMs used in this work. We select books from the romance genre as they frequently use tropes (*e.g.*, **enemies-to-lovers, friends-to-lovers, second chance, and forbidden love**) with evolving relationships between the main characters to make the story interesting. We manually refer to online reading forums such as Goodreads[7] to obtain the specific trope depicted in the selected books and their main characters. We use this information for global-level evaluation of the predicted trajectory for a book (§8). We perform experiments for a pair of main characters per book however, the proposed strategies are agnostic to the pair of characters and can be used for any two characters in theory.

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

[6] https://www.sparknotes.com/lit/

[7] https://www.goodreads.com/

5.2 Preprocessing Books

We preprocess books using BookNLP (Bamman et al., 2014) library[8] to get coreferences for characters in a book. We first divide the book text into non-overlapping passages of human-readable length ($100 - 200$ words). Then, replace the first occurrence of any third-person pronouns used as subject with a representative alias for a character. The most frequently used proper noun for a character is considered the representative alias for that character. We do such a replacement to ensure the comprehensibility of a standalone passage. We refer to the above process as **coreference substitution**. We obtain 644 ± 104 passages per book, of which 98 ± 128 passages have both main characters mentioned in them. Huge variation is due to differing author writing styles. We do not perform any coreference substitution for the complete passages strategy since prior passages are provided as-is and as per centering theory coreferences are used to maintain local coherence (Grosz et al., 1995).

5.3 Summarizer and Predictor Models

We use open-sourced LLMs from three families, namely, Llama3.1-8B-chat (Dubey et al., 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Gemma2-9B (Team et al., 2024), to obtain the iterative summaries and predict relationship and evolution type for the *Summary with Passage* and *Complete Summary* strategies. However, for the *Complete Passages* strategy, we use Llama3.1-8B-chat with a maximum of $30K$ context window size.

6 Evaluation Without Gold Labels

One of the major challenges of tracking evolution in the relationship between characters is the unavailability of gold labels and the difficulty in collecting crowd-sourced annotations due to the length of the books; making it extremely expensive, and cognitively challenging. We make a novel contribution by providing insights into the feasibility of this task without gold labels by analyzing the agreement between predictions from multiple LLMs. Additionally, we conduct a quantitative (§9.1) and qualitative (§9.2) manual analysis of the predictions.

Owing to the increasing use of LLM-as-evaluators (Chan et al.; Gu et al., 2024) and LLM-as-annotators (Chiang and Lee, 2023; Tan et al., 2024), we hypothesize that if multiple LLMs agree on a label then it is more likely to be the gold

[8] https://github.com/booknlp/booknlp

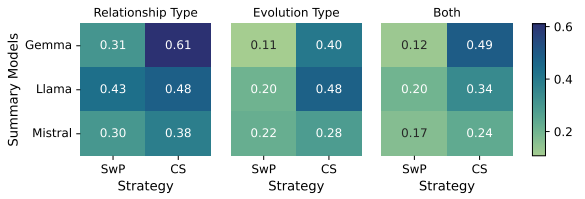


Figure 3: Krippendorff's alpha between different predictions for *Summary with Passage* (SwP) and *Complete Summary* (CS) strategies using summaries generated from various summary models.

label (Liang et al., 2024; Chern et al., 2024). Therefore, we thoroughly analyze the quality of the consensus predictions from multiple LLMs (conditioned on the input) via Krippendorff's alpha (α), and compare it against the agreement between human annotators. We also use α scores to compare agreement between different strategies (Table 3 and Figure 4), and preprocessing methods (Figure 5 and Figure 6). We manually examine the consensus predictions at a global- and local-level to provide an upper bound on the performance of LLMs for this task. **Global-level evaluation** measures the accuracy of correctly predicting the trope given the predicted trajectory of evolution in the relationship between two characters for a book. For **local-level evaluation**, a subset of examples per book per trope is manually annotated and compared against the consensus predictions. We report scores separately for **relationship** and **evolution type** as well as when **both** are considered together.

7 Findings

Agreement between prediction models. We report the agreement (α scores) between predictions from different LLMs for *Summary with Passage* and *Complete Summary* strategies that use summaries generated from various summary models in Figure 3. While this analysis depends on the correctness of the summaries generated from LLMs, we do not explicitly evaluate their correctness as LLM-generated summaries have been shown to be on par with human-written summaries (Goyal et al., 2022; Zhang et al., 2024). Low-to-moderate agreement suggests that tracking evolution in the relationship is a challenging task. Lower agreement for evolution type than relationship type emphasizes the difficulty of predicting fine-grained evolution. We observe higher agreement between predictions for *Complete Summary* than *Summary with Passage* as summaries contain more direct evidence of the

Strategy	Relationship Type	Evolution Type	Both
Summary with Passage	0.24	0.20	0.13
Complete Summary	0.22	0.08	0.11

Table 2: Agreement between predictions across summary and prediction models.

Strategy	Relationship Type	Evolution Type	Both
Summary with Passage	0.50	0.44	0.38
Complete Summary	0.47	0.12	0.17

Table 3: Agreement between predictions from *Complete Passages* and consensus predictions for *Summary with Passage* and *Complete Summary* strategies.

Strategy	Relationship Type (α)	Evolution Type (α)	Both (α)
Summary with Passage	86.04 (0.70)	44.18 (0.26)	39.54 (0.31)
Complete Summary	88.37 (0.79)	70.93 (0.58)	67.44 (0.61)

Table 4: **Local-level evaluation:** Accuracy of consensus predictions averaged across annotations from humans. Scores in the parenthesis denote the agreement (α) between human annotators.

type and nature of the relationship whereas in the presence of a more granular passage, the evidence needs to be inferred. Interestingly, agreement between predictions across all the summary models is higher for *Summary with Passage* than the *Complete Summary* strategy (see Table 2). This indicates that using a passage acts as a regularizer for mitigating the effect of differences in summaries generated from different summarizers.

***Summary with Passage* is a better approximation of *Complete Passages* than the *Complete Summary* strategy.** To analyze which strategy – *Summary with Passage* or *Complete Summary* – using a short context window best approximates *Complete Passages* that uses a long context window, we report agreement between the predictions from *Complete Passages* strategy and the consensus predictions across all the summarizers and predictors for the two shorter-window strategies in Table 3. Combining a passage with a prior summary strikes a good balance in providing the information necessary for the task, compared to using the complete summary. This supports the regularization aspect of using a passage, as seen in Table 2.

8 Manual Evaluation

How accurate are the consensus predictions from LLMs? To study the upper bound performance achievable from the proposed strategies, two annotators label the relationship and evolution type

Strategy	Summarizer	Accuracy (%)
Majority Trope	N/A	54.54
Complete Passages	N/A	63.64
Summary with Passage	Llama	27.27
	Mistral	18.18
	Gemma	18.18
	Overall	21.21
Complete Summary	Llama	9.09
	Mistral	9.09
	Gemma	0.00
	Overall	6.06

Table 5: **Global-level evaluation**: Percentage accuracy for predicting the trope of a book (by human annotators) from the predicted trajectory from various strategies.

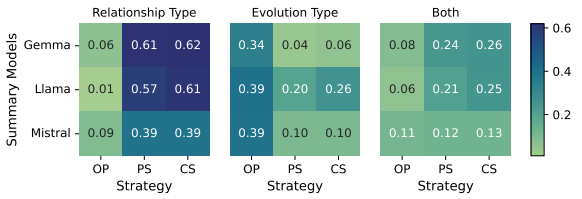


Figure 4: Krippendorff's alpha between consensus predictions from *Summary with Passage* and *Only Passage* (OP) or *Previous Summary* (PS) ablations. Agreement with *Complete Summary* (CS) is shown to emphasize its difference as opposed to *Previous Summary*.

conditioned on the input for different strategies. For local-level evaluation, we first select one book per trope (4 books in total) that attains the highest prediction agreement over all the predictors and summarizers across different strategies. Then, we sample a maximum of 20 passages (where both the characters are mentioned) per selected book and consensus predictions from the summarizer with the highest agreement[9] for each strategy. Passages are selected at random such that the consensus predictions from different strategies are different as it has a two-fold benefit: (1) the accuracy of prediction for the sampled passages acts as a good approximation of overall accuracy as the two strategies will have the same accuracy for the same predictions, and (2) it makes it easier to qualitatively compare the two strategies (as shown in Table 7). We report the accuracy of prediction averaged over the two annotators and α (in parenthesis) between them in Table 4. We observe that agreement entails accuracy for both strategies. A similar agreement between humans, as observed for LLMs (see Figure 3), indicates that while relationship type can be determined with high agreement, evolution type prediction is challenging for humans as well.

How accurately can trope be identified from the LLM predicted trajectory? For global-level evaluation, we present a visualization (similar to Figure 1) of the trajectory of evolution in the relationship between two characters to annotators[10] and ask them to select the best applicable trope out of enemies-to-lovers, friends-to-lovers, second-chance, and forbidden love[11]. We keep the book and characters' names anonymous to the annotators to ensure no use of online resources. We provide visuals obtained from the predictor having the highest agreement with the consensus predictions for each strategy and summarizer[12]. As mentioned in §5.1, we have gold trope labels for each book to compute the accuracy of trope prediction.

Table 5 shows that trope can be predicted with the highest accuracy when all the passages are provided to an LLM with a large context window (*i.e. Complete Passages*). Higher accuracy for summaries from Llama indicates that it has a better understanding of social relationships than Gemma or Mistral. While we see a higher local-level accuracy for *Complete Summary* than *Summary with Passage* strategy (Table 4), we observe a reverse trend for trope prediction. This shows that an overall summary is unable to capture the fine-grained details; aligning with our previous finding that *Summary with Passage* is a better approximation of *Complete Passages* than *Complete Summary* (see Table 2). Lower accuracy than a majority baseline emphasizes the difficulty of this task when information is not provided at the highest granularity.

9 Analysis and Discussion

We present an ablation study of *Summary with Passage* strategy, need for coreference substitutions, and intermediate passages in §9.1 and shed light on the challenges with this task in §9.2.

9.1 Quantitative Analysis

Evolution type is determined (mostly) based on the passage whereas relationship type is (majorly) influenced by the previous summary. To

[9] Summaries from Llama resulted in higher agreement between predictions as compared to that from Gemma or Mistral.

[10] Manual examination is done internally by people who frequently read novels.

[11] We also provide "cannot be determined" and "others" as options.

[12] 11 visualizations per strategy per summarizer.

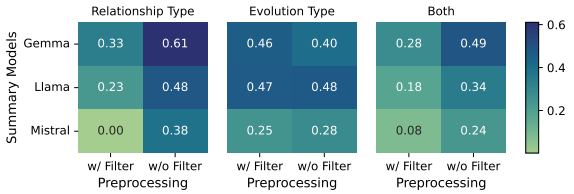
Strategy	Relationship Type	Evolution Type	Both
Only Passage	59.43 (0.34)	49.05 (0.33)	39.62 (0.23)
Previous Summary	82.95 (0.59)	53.41 (0.40)	45.45 (0.32)
Summary with Passage	72.64 (0.72)	41.51 (0.24)	33.02 (0.38)

Table 6: Accuracy (%) of consensus predictions averaged across annotations from human annotators. Scores in the parenthesis denote the agreement (α) between human annotators.

analyze the source of information used to predict the evolution and relationship types in the *Summary with Passage* strategy, we perform an ablation study by using only the passage or the previous summary and compare the consensus predictions with that from using both the passage and the previous summary. Higher agreement (Figure 4) between the *Previous Summary* and *Summary with Passage* strategy than *Only Passage* for relationship type shows that LLMs rely on information in the summary for relationship type prediction. However, the evolution type predictions are determined based on the provided passage. As expected, agreement between predictions from the complete summary is higher than that from the previous summary since it contains more information. Additionally, we ask two annotators to label a subset of examples selected in the same way as in §8 for local-level evaluation and report the results in Table 6. Low agreement between annotators (in parenthesis) except for relationship predictions from *Summary with Passage* strategy shows that this task is difficult even for humans. This is due to the involved subjectivity leading to different annotations (see Table 8 for examples). Accuracy for relationship predictions for *Only Passage* is much lower than that for *Summary with Passage* due to the possibility of multiple interpretations when a passage is provided out-of-context. However, evolution prediction is less accurate when a previous summary is provided with a passage. This may happen when relationships are in a transition phase, or characters may have different emotional states toward each other. Since a summary captures all this information, it may be difficult to infer an evolution type with certainty. We discuss such examples in §9.2.

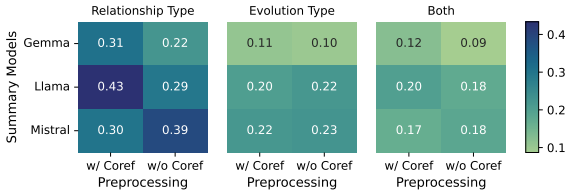
Intermediate context is a useful source of information. Prior studies (Chaturvedi et al., 2016; Iyyer et al., 2016; Chaturvedi et al., 2017) that model evolution in relationships have focused solely on passages where both characters are mentioned. In contrast, we hypothesize that interactions

(a) Summary with Passage

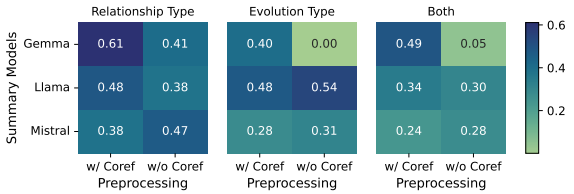


(b) Complete Summary

Figure 5: Krippendorff's alpha between consensus predictions from *Summary with Passage* and *Complete Summary* which use summaries generated from passages with (w/ Filter) and without filtering (w/o Filter) the passages that do not mention both the characters.



(a) Summary with Passage



(b) Complete Summary

Figure 6: Krippendorff's alpha between consensus predictions from *Summary with Passage* and *Complete Summary* with (w/ Coref) and without substituting the character coreferences (w/o Coref) in the passages.

between other characters or one of the main characters with others is a useful source of information for this task. To test this, we employ the same strategies as described in §4 but only use the passages where both the characters are mentioned and compare the obtained agreements with those when passages without both character mentions are not filtered. Figure 5 indeed shows that intermediate context results in higher agreement for relationship predictions when passages are not filtered than when filtered. However, we do not observe significant improvement for evolution prediction.

Coreference resolution results in (mostly) higher or comparable agreement between predictions than without it. We apply the *Summary with Passage* and *Complete Summary* strategy on passages from a book without coreference substitutions (described in §5.2) to analyze its impact on the agreement between predictions as shown in Figure 6. While we mostly see a higher agreement between predictions when coreference substitution is done during data preprocessing (w/ Coref) than in its absence (w/o Coref), we also see instances of lower or comparable agreement. Manual analysis reveals that in the absence of a specific character mention, LLMs tend to assume that the pronouns refer to the characters understudy both during summary generation, and relationship and evolution prediction. This is a result of the widely acknowledged issue of hallucination (Ye et al., 2023; Huang et al., 2023) and context sensitivity (Min et al., 2022) in LLMs.

9.2 Qualitative Analysis

Maintaining a running summary helps resolve the ambiguity between different relationship types. Manual analysis reveals that providing a previous summary helps propagate the prior knowledge about the relationship that can make predictions more certain, and resolve any ambiguity due to insufficient information or out-of-context passages (see Table 7 in appendix for examples).

Uncertainty in relationship or evolution type prediction results in disagreements between humans. We find that humans might have different interpretations when a relationship is in a "transition/developing" phase, the two characters have different emotional states towards each other, or a phrase with multiple interpretations is mentioned in the text, leading to different annotations (see Table 8 in appendix for examples).

Failure cases. Analysis in Table 9 (in appendix) shows that LLMs rely on surface-level cues, tend to resolve pronouns to the character in question in the absence of an explicit mention, and are sensitive to subtle changes in the context (such as substituting pronouns for other characters in the context). Such behavior raises questions on the *true* understanding of evolving social relationships in LLMs and if they are right for the wrong reasons.

10 Related Work

Existing works that examine relationships between characters in narratives either use a fixed set of coarse-grained relations, such as cooperative or non-cooperative (Srivastava et al., 2016; Chaturvedi et al., 2016) and familial or professional (Makazhanov et al., 2014; Massey et al., 2015; Azab et al., 2019) or learn a set of relationship descriptors (Iyyer et al., 2016). Others classify emotional relationships between characters in fanfiction stories (Kim and Klinger, 2019b) and harry potter novel (Zehe et al., 2020) following Kim and Klinger (2019a). Another line of research analyzes the polarity and intensity of emotions of characters towards each other (Nalisnick and Baird, 2013) in Shakespearean plays, or classifies interpersonal relationships from dialogues in TV series (Chen et al., 2020), movie scripts (Jia et al., 2021) or detective narratives (Zhao et al., 2024).

While the above works consider static relationships, Chaturvedi et al. (2016) model the evolution of interpersonal relationships in novels in a supervised setting, requiring manual annotations, and model relationships as binary polarities. Whereas Iyyer et al. (2016) introduce an unsupervised method, RMN (Relationship Modeling Network), to model evolving relationships by learning a sequence of discrete states depicting the relationship between the two characters. Qamar et al. (2021) employ psychological models to classify movie dialogues into attachment styles and association types to analyze the transformation between relationships. However, we focus on its evolution.

11 Conclusion

This work tracks the evolution in the relationship between characters in books by proposing several strategies that differ in the granularity of information provided to the LLMs to assess their understanding of social relationships. Without gold annotations, our analysis of agreement between predictions from multiple LLMs shows that providing a running summary of the type and nature of evolution in the relationship between the characters along with a passage is a better approximation of a strategy that uses all the passages until a point than providing a complete summary. Overall, low-to-moderate agreement between LLMs as well as between humans shows the difficulty of the task. While human disagreement can be attributed to their differing interpretations of the context, qualitative analysis reveals that LLMs adopt surface-level cues, and are sensitive to subtle changes in the provided context raising questions on their *true* understanding of social relationships.

Limitations

We acknowledge the below limitations of this work.

Linear plot structure assumption We assume linear plot structure of the books in this work to assess how LLMs perform in a straightforward setting. However, plot structure can be nonlinear and complex such as worlds within worlds wherein the narrative timelines and chronological timelines could be different. We leave tracking evolution in relationships in such books for future research.

Coverage of relationship and asymmetric evolution We use a subset of relationship types that commonly occur in books from the romance genre between main characters. However, we acknowledge that the set of relationships is not exhaustive and may need to be updated based on the genre of books used and the type of relationships that occur in such books. As shown in the qualitative examples, evolution in relationship may be different from each character's perspective. We leave study of such asymmetric evolution in relationships for future work.

Potential errors in conference substitutions Coreference resolution at book-length is still an open problem in NLP (Toshniwal et al., 2020; Xia and Van Durme, 2022; Guo et al., 2023). While we use widely known BookNLP (Bamman et al., 2014) toolkit, we believe that incorrect coreferences could result in misinterpretation of the text and lead to prediction errors. Future work may further investigate the impact of incorrect coreference substitutions on the task of tracking evolution in relationships.

Input conditional evaluation of strategies As gold annotations are not available for this task due to the length of the books and the cognitively challenging nature of the task, our agreement analysis as well as local-level manual evaluation of predictions is conditioned on the input of the specific strategy used. However, we believe that the global-level evaluation via trope prediction provides a good upper bound on the performance achievable from different strategies for this task. An ideal scenario would be when the predictions from different strategies are compared to gold annotations available at different points in the book.

Potential errors in the LLM generated summaries While LLM-generated summaries are on par with human-written summaries (Goyal et al., 2022; Zhang et al., 2024), we acknowledge that summaries may be prone to incoherence and factual inconsistencies. This is potentially the reason behind lower performance for summaries from Gemma and Mistral models. Since the focus of this work was on tracking evolution in relationships, we leave further analysis of summaries until each passage of the book and its impact on the performance of this task for future research.

Ethical Considerations

Our study presents a systematic approach for evaluating LLMs for their social reasoning capabilities and hence does not inherently pose direct risks. However, it is important to emphasize that predictions from LLMs may be influenced by inherent biases that may get ingrained in them during the pretraining stage. Therefore, before deploying the proposed strategies in our work, the predictions should be human-evaluated and debiased to ensure safety and avoid any potential social harm. The dataset used in this work was acquired by directly contacting the authors of that paper. Due to copyright issues, the dataset is not publicly available and we make sure that the data is handled properly with no redistribution.

References

Mahmoud Azab, Noriyuki Kojima, Jia Deng, and Rada Mihalcea. 2019. Representing movie characters in dialogues. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 99–109, Hong Kong, China. Association for Computational Linguistics.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

Philip Blumstein and Peter Kollock. 1988. Personal relationships. *Annual Review of Sociology*, 14(1):467–490.

Gordon H Bower and Daniel G Morrow. 1990. Mental models in narrative comprehension. *Science*, 247(4938):44–48.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Orson Scott Card. 1999. Characters & viewpoint: Elements of fiction writing. *Cincinnati, OH: Writer's Digest Books*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023a. Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv preprint arXiv:2305.00118*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023b. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 610–614, Marseille, France. European Language Resources Association.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Gregory Currie. 2009. Narrative and the psychology of character. *The journal of aesthetics and art criticism*, 67(1):61–71.

Sebastian Deri, Jeremie Rappaz, Luca Maria Aiello, and Daniele Quercia. 2018. Coloring in the links: Capturing social ties as they are perceived. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Susan T Fiske, Shelley E Taylor, Nancy L Etcoff, and Jessica K Laufer. 1979. Imaging, empathy, and causal attribution. *Journal of Experimental Social Psychology*, 15(4):356–377.

Morton Ann Gernsbacher, Brenda M Hallada, and Rachel RW Robertson. 1998. How automatically do readers infer fictional characters' emotional states? *Scientific studies of reading*, 2(3):271–300.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Dual cache for long document neural coreference resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15272–15285.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Qi Jia, Hongru Huang, and Kenny Q Zhu. 2021. Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13125–13133.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

David Jurgens, Agrima Seth, Jackson Sargent, Athena Aghighi, and Michael Geraci. 2023. Your spouse needs professional help: Determining the contextual appropriateness of messages through modeling social relationships. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10994–11013, Toronto, Canada. Association for Computational Linguistics.

XJ Kennedy and Dana Gioia. 1983. Literature: An introduction to fiction. *Poetry, Drama, and writing*.

XJ Kennedy, Dana Gioia, and Dan Stone. 2013. *Literature: An introduction to fiction, poetry, drama, and writing*. Pearson.

Evgeny Kim and Roman Klinger. 2019a. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2019b. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Gabrielle Lissauer. 2014. *The Tropes of Fantasy Fiction*. McFarland.

Aibek Makazhanov, Denilson Barbosa, and Grzegorz Kondrak. 2014. Extracting family relationship networks from novels. *arXiv preprint arXiv:1405.0603*.

Philip Massey, Patrick Xia, David Bamman, and Noah A Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint arXiv:1512.00728*.

Robert McKee. 1997. *Story: style, structure, substance, and the principles of screenwriting*. Harper Collins.

Gerald Mead. 1990. The representation of fictional character. *Style*, pages 440–452.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric T. Nalisnick and Henry S. Baird. 2013. Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, Sofia, Bulgaria. Association for Computational Linguistics.

Alison H Paris and Scott G Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Saira Qamar, Hasan Mujtaba, Hammad Majeed, and Mirza Omer Beg. 2021. Relationship identification between conversational agents using emotion analysis. *Cognitive Computation*, 13:673–687.

Farzana Rashid and Eduardo Blanco. 2018. Characterizing interactions and relationships between people. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4404, Brussels, Belgium. Association for Computational Linguistics.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Bryan Stroube. 2003. Literary freedom: Project gutenberg. *XRDS: Crossroads, The ACM Magazine for Students*, 10(1):3–3.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models

based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2021. PRIDE: Predicting Relationships in Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4636–4650, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to ignore: Long document coreference with bounded memory neural networks. *arXiv preprint arXiv:2010.02807*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. *arXiv preprint arXiv:1902.01390*.

Myron Wish, Morton Deutsch, and Susan J Kaplan. 1976. Perceived dimensions of interpersonal relations. *Journal of Personality and social Psychology*, 33(4):409.

Patrick Xia and Benjamin Van Durme. 2022. Online neural coreference resolution with rollback. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 13–21.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Albin Zehe, Julia Arns, Lena Hettinger, and Andreas Hotho. 2020. Harrymotions-classifying relationships in harry potter based on emotion analysis. In *SwissText/KONVENS*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Runcong Zhao, Qinglin Zhu, Hainiu Xu, Jiazheng Li, Yuxiang Zhou, Yulan He, and Lin Gui. 2024. Large language models fall short: Understanding complex relationships in detective narratives. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7618–7638, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Detailed Experimental Setup

A.1 Implementation Details

We generate summaries of a maximum 300 words given passages of a maximum 200 words. We use nucleus sampling (Holtzman et al., 2019) with radius of $p = 0.9$ and top $k = 50$ tokens for generating summaries and greedy decoding for the prediction tasks.

A.2 Zero-shot Prompt for Obtaining the Summaries and Predictions

We use different prompts to obtain the summary of the first passage (Prompt A.1) and to update the previous summary with a new passage (Prompt A.2). For compressing the summary within a word limit, we use Prompt A.3).

We use the Prompt A.4, Prompt A.6, and Prompt A.5 to get the relationship and evolution type predictions from an LLM for the *Complete Passages*, *Complete Summary*, and *Summary with Passage* strategies, respectively. Iteratively asking for relationship type and then evolution type helps in presenting only the required question and information to an LLM and makes it easier to parse the output.

A.3 Manual Evaluation Details

The manual evaluation was done by experienced annotators who read novels. Two of them are doctoral students and the other two have undergraduate degrees. We clearly explain the annotation task and run a small pilot followed by a discussion to ensure the task annotation is clear.

Prompt A.1: Iterative Summary Generation: First Passage

```
System Prompt: You are a helpful assistant who
follows the instructions. No preambles and
postambles. Avoid explanations if not asked
explicitly.

Prompt: Below is the beginning part of a
story from a book:

—
{story}
—

We are going over segments of a story
sequentially to gradually update one
comprehensive summary depicting the evolution
in the relationship between {char_a} and
{char_b}. Write a summary of the evolution
in the relationship between {char_a} and
{char_b} as the story progresses. Make sure
to include vital information related to key
events that shape the relationship between
{char_a} and {char_b}, their objectives, and
motivations. The story may feature non-linear
narratives, flashbacks, switches between
alternate worlds or viewpoints, etc. Therefore,
you should organize the summary so it presents
a consistent and chronological narrative.
Despite this step-by-step process of updating
the summary, you need to create a summary that
seems as though it is written in one go. The
summary should roughly contain 300 words.

Constraint: If the provided segment does
not mention both {char_a} and {char_b}, then
do not make up or predict anything regarding
the relationship between the characters. Just
provide a general summary of the provided
segment or keep it empty.

Summary:
```

Prompt A.2: Iterative Summary Generation: Updating Previous Summary

```
System Prompt: You are a helpful assistant who follows the instructions. No preambles and postambles. Avoid explanations if not asked explicitly.

Prompt: Below is a segment of a story from a book:

——
{story}
——

Below is a summary of the evolution in the relationship between {char_a} and {char_b} up until this point in the story.

——
{summary}
——

We are going over segments of a story sequentially to gradually update one comprehensive summary depicting the evolution in the relationship between {char_a} and {char_b}. You are required to update the provided summary to incorporate any new vital information related to the relationship between {char_a} and {char_b} that is present in the current segment of the story. This information may relate to key events, turning points in the relationship between the characters, their objectives, and motivations. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative. Despite this step-by-step process of updating the summary, you need to create a summary that seems as though it is written in one go. The updated summary should roughly contain 300 words.

Constraint: If the provided segment does not mention both {char_a} and {char_b}, then avoid making up or predicting anything regarding the relationship between the characters. Just copy the provided summary as-is or update it with the general aspects of the story or keep it empty.

Updated summary:
```

Prompt A.3: Compress Summary

```
System Prompt: You are a helpful assistant who follows the instructions. No preambles and postambles. Avoid explanations if not asked explicitly.

Prompt: Below is a summary of the relationship between {char_a} and {char_b} from a part of a story:

—
{Summary}
—

Currently, this summary contains {summary_length} words. Your task is to condense it to less than 300 words while maintaining the chronological order. The condensed summary should remain clear, overarching, and fluid while being brief. Whenever feasible, maintain details about key events that shape the relationship between {char_a} and {char_b}, how does the relationship evolve over time, character's objectives, and motivations - but express these elements more succinctly. Remove insignificant details that do not add much to the overall evolution in the relationship between {char_a} and {char_b} and phrases like "in this .. segment", "in this part ... story", etc. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative.

Condensed summary (to be within 300 words):
```

Prompt A.4: Relationship and Evolution Type Prediction Prompt for Complete Passages

```
System Prompt: You are a helpful assistant who follows the instructions. No preambles and postambles. Avoid explanations if not asked explicitly.

Prompt: Based on the provided context and the following segment, answer the below questions about the type of relationship between {char_a} and {char_b} and its evolution by the end of the provided segment.

Context:
——
{previous passages}
——

Segment:
——
{segment}
——

Are {char_a} and {char_b} mentioned in the provided segment? Answer in one word <ANS> ["yes", "no", "unsure"] </ANS>.
{Model's output}

Can you infer any type of relationship between {char_a} and {char_b} from the segment? Answer in one word <ANS> ["yes", "no", "unsure"] </ANS>
{Model's output}

Choose the type of relationship between {char_a} and {char_b} from these options: acquaintances, strangers, friends, best friends, romantic interest, dating, engaged, married, separated, divorced, enemies, spouse, ex-spouse, one-sided romantic interest, ex-romantic interest, others or cannot be determined. Answer only from the provided options. Relationship type:
{Model's output}

Is the chosen relationship between {char_a} and {char_b} evolving "positively", "negatively", is "stable" or "nothing can be determined" by the end of the segment? A "positive" evolution can result from deepening connection, increasing trust, support or respect, spending more time together etc. A "negative" evolution means any tension or straining relationship that can result from conflicts, arguments, distrust, disrespect, lack of support, or misunderstandings. A "stable" relationship means there is neither positive nor negative evolution. Do not provide any explanation. Evolution type:
```

Prompt A.5: Relationship and Evolution Type Prediction Prompt for Summary with Passage

```
System Prompt: You are a helpful assistant who follows the instructions. No preambles and postambles. Avoid explanations if not asked explicitly.

Prompt: Based on the summary of the evolution in type and nature of the relationship between {char_a} and {char_b} until a point in a book and the following segment answer the below questions about the type of relationship between {char_a} and {char_b} and its evolution by the end of provided segment.

Summary:
——
{summary}
——

Segment:
——
{segment}
——

Are {char_a} and {char_b} mentioned in the provided segment? Answer in one word <ANS> ["yes", "no", "unsure"] </ANS>.
{Model's output}

Can you infer any type of relationship between {char_a} and {char_b} from the segment? Answer in one word <ANS> ["yes", "no", "unsure"] </ANS>
{Model's output}

Choose the type of relationship between {char_a} and {char_b} from these options: acquaintances, strangers, friends, best friends, romantic interest, dating, engaged, married, separated, divorced, enemies, spouse, ex-spouse, one-sided romantic interest, ex-romantic interest, others or cannot be determined. Answer only from the provided options. Relationship type:
{Model's output}

Is the chosen relationship between {char_a} and {char_b} evolving "positively", "negatively", is "stable" or "nothing can be determined" from the segment? A "positive" evolution can result from deepening connection, increasing trust, support or respect, spending more time together etc. A "negative" evolution means any tension or straining relationship that can result from conflicts, arguments, distrust, disrespect, lack of support, or misunderstandings. A "stable" relationship means there is neither positive nor negative evolution. Do not provide any explanation. Evolution type:
```

Prompt A.6: Relationship and Evolution Type Prediction Prompt for Complete Summary

```
System Prompt: You are a helpful assistant who follows the instructions. No preambles and postambles. Avoid explanations if not asked explicitly.

Prompt: Based on the given summary that depicts the evolution in the relationship between char_a and char_b until a point in a book answer the below questions about the type of relationship between {char_a} and {char_b} and its evolution at the end of the summary.

Summary:
--
{summary}
--

Are {char_a} and {char_b} mentioned in the provided summary? Answer in one word <ANS> ["yes", "no", "unsure"] </ANS>.
{Model's output}

Can you infer any type of relationship between {char_a} and {char_b} from the summary? Answer in one word <ANS> ["yes", "no", "unsure"] </ANS>
{Model's output}

Choose the type of relationship between {char_a} and {char_b} from these options: acquaintances, strangers, friends, best friends, romantic interest, dating, engaged, married, separated, divorced, enemies, spouse, ex-spouse, one-sided romantic interest, ex-romantic interest, others or cannot be determined. Answer only from the provided options. Relationship type:
{Model's output}

Is the chosen relationship between {char_a} and {char_b} evolving "positively", "negatively", is "stable" or "nothing can be determined" by the end of the summary? A "positive" evolution can result from deepening connection, increasing trust, support or respect, spending more time together etc. A "negative" evolution means any tension or straining relationship that can result from conflicts, arguments, distrust, disrespect, lack of support, or misunderstandings. A "stable" relationship means there is neither positive nor negative evolution. Do not provide any explanation. Evolution type:
```

B Qualitative and Quantitative Analysis

Summary: *Mina stands at a crossroads, torn between her village and the allure of the open sea, driven by her love for Shin. Her memories reveal the early stages of their romance, suggesting her love for him may be a choice rather than a predetermined fate. As Mina's devotion to Shin reaches a boiling point, she breaks the Sea God's three rules, and Joon's concern for her safety demonstrates the strong bond between them. . . . She encounters Shin, who looks at her with longing in his eyes, breaking her heart. Shin's words of encouragement, " "Don't chase fate, Mina. Let fate chase you," remind her to find her own path and destiny. Mina confesses her feelings to Shin, and he reciprocates, stating that he doesn't need the Red String of Fate to know that he loves her. They share a passionate kiss, and Shin says, "Lord Crane was mistaken. He said once the Red String of Fate was formed, you would know how to break the curse."*
Passage: *Namgi says. He leans back, and I get a good look at his face. There's joy there, and wonder. "We know everything, about the emperor, about the Sea God. Shin is the Sea God! Can you believe it?" " Where is he?" I ask. "In the hall. We arrived right before you." Kirin approaches from behind Namgi, his always astute eyes watching me carefully. "What were you saying, Mina? That you wouldn't see us before...?" I release Namgi, stepping back.*
Predictions: cannot be determined, romantic interest

Summary: *Jana's relationship with Anil began with a transformative experience, marked by a week and a half of intense physical connection. . . . As they traveled together, Jana and Anil grew closer, visiting Tajikistan . . . In London, they transitioned from traveling companions to intimate partners, engaging in a hard-and-fast fling amidst their days of attending meetings and nights in a tiny hotel room. . . . Their connection deepened, and they found themselves lost in intimate moments, discussing their work in the development field and goals of bringing grassroots-style microdevelopment to a larger scale. dots Jana's trust in Anil was shaken when she received anonymous messages accusing her of sleeping with a married man, claiming Anil was still married to the sender's sister. . . .*
Passage: *Jana knew she was falling in love with him. And maybe love was a little blind. She picked up her phone again, not to call Anil, but to message Rasheed, the manager of the project in Tajikistan. The one who Anil had been visiting when this relationship started. Jana: Rasheed, is Anil married? It was the middle of the night in Tajikistan. Jana wasn't expecting an answer. But he responded. Rasheed: Have you asked him that question?*
Predictions: one-sided romantic interest, romantic interest

Table 7: Sample relationship predictions depicting importance of using a previous summary which helps resolve uncertain predictions, and propagate prior context to avoid misinterpretation from just the passage. Color denotes the predictions and evidences from *Only passage* and *Summary with Passage* strategy.

Summary: *Jana's relationship with Anil began with a transformative experience, marked by a week and a half of intense physical connection. . . . However, Jana's trust in Anil was shaken when she received anonymous messages accusing her of sleeping with a married man, claiming Anil was still married. Anil's revelation that he was indeed married marked the end of their relationship, and Jana became pregnant with his child. Years later, . . . Anil moved to Toronto to be close to Imani, and as co-parents, Jana and Anil discussed their complicated family dynamics. Jana expressed concerns about becoming overprotective, while Anil demonstrated a willingness to be involved in Imani's life. Jana struggled with her past and her growing closeness to Anil, confronting him about his past behavior and accusing him of trying to stroke her ego. . . . As they reconnect, Jana questions whether she is letting other people's opinions get in the way of her own happiness, particularly in regards to her relationship with Anil. Jana is starting to inch out of her comfort zone, and now considers a date with Anil, suggesting a possible rekindling of their relationship.*
Passage: *Did you forget what happened at Hatari? Jana asked. Anil chuckled low, sending a shiver down Jana's spine. . . . "I've honestly thought of little else," he said. And there it was. He'd been as preoccupied with thoughts of that night as she'd been. Jana could feel a heat burning inside her. He still wanted her.*
Discussion: Jana and Anil are ex-romantic partners who are co-parenting their daughter and are considering to give another chance (romantic interest) to their relationship.

Summary: *Jana's relationship with Anil began with a transformative experience, marked by a week and a half of intense physical connection. . . . However, Jana's trust in Anil was shaken when she received anonymous messages accusing her of sleeping with a married man, claiming Anil was still married. Anil's revelation that he was indeed married marked the end of their relationship. Jana became pregnant with his child, and they navigated a co-parenting agreement with the help of a family lawyer. Years later, Anil surprised Jana . . . revealing he wanted to surprise their daughter Imani . . . This gesture suggested a desire to reconnect with Jana and their daughter. . . . Their recent interactions highlighted the challenges they still face, including Anil's condescending behavior and Jana's lingering discomfort. . . . Jana's hesitation stems from her need for self-care, as being around Anil triggers memories and makes it difficult for her to think straight.*
Passage: *Jana couldn't travel for two weeks alone with just him and Imani. His betrayal still hurt too much. She could finally take a trip with Imani, and this would taint it. "Come on, Jana," Anil said. "I don't want to break your mother's heart, either. She's so great for Imani."*
Discussion: Jana and Anil are ex-romantic partners as well as co-parents. While Anil is putting in efforts to reconnect with Jana and his daughter, Jana has unresolved emotions and is hesitant resulting in a positive evolution in the relationship from Anil's side however, still negative from Jana's perspective.

Passage: *Kirin strides in, bowing low. His keen eyes glance at Shin's hand, still holding my own. "You called for me?" "Mina's been hurt." "Ah, I see." I frown at the two of them, the unspoken words thick in the air. Why had Shin asked for Kirin and not a physician? As Shin releases my hand, Kirin reaches inside his robes and pulls out a small silver dagger. . . . I only have a moment to gape before he grabs my wrist, placing his now bloodied hand over my burned one.*
Dicussion: Here, "holding hands" was interpreted as showing care as *romantic interests* by one annotator while from another's perspective it was considered as a gesture any *friend* would do if someone is injured.

Passage: *It was the cutest thing Jana had ever seen. She lifted Imani in her arms to see it better. Everyone in the Land Cruiser was in awe, giddy with excitement. As the drive continued, they saw gazelles, the most vibrant striped zebras yet, and some giraffes. But Jana understood why this was called the elephant park. There were so many elephants. On their own, in herds, at the watering hole ,Äî everywhere. Anil kept pointing out new ones, and Imani eventually stopped counting (she really couldn't get past forty, anyway).*
Discussion: As Jana and Anil are spending time together one annotator considered it as a positive evolution however, for another, there was not enough information to determine evolution type from the passage.

Summary: *Jana's relationship with Anil began with a transformative experience, marked by a week and a half of intense physical connection. . . . Their connection deepened in London, . . . However, Jana's trust in Anil was shaken when she received anonymous messages accusing her of sleeping with a married man, claiming Anil was still married. Anil's revelation that he was indeed married marked the end of their relationship. Jana became pregnant with his child, and they navigated a co-parenting agreement . . . This gesture suggested a desire to reconnect with Jana and their daughter. . . . Now, Anil is considering Jana for a director of research and programs position, creating a delicate situation for Jana. She must navigate her past anger and work with Anil to co-parent their daughter effectively.*
Passage: *And now she had two weeks with Anil, this was the perfect time to put it all behind her for Imani's well-being and her own. It was time to show everyone, including Anil Malek, that the last five years hadn't broken Jana. Most of all, it was time for Jana to show herself that. . . . It was a revised wedding schedule. Jana assumed Elsie had rearranged everything so Dr. Lopez wasn't in the same activities as Mom, Jana, Anil, or Imani.*
Discussion: While the evolution is negative as per one annotator due to 'creating a delicate situation for Jana', nothing 'can be determined' from another's perspective as not there is no direct interaction between Anil and Jana in the passage but it mostly mentions Jana's feelings.

Table 8: Examples where relationship and evolution prediction may be uncertain and open to interpretation leading to disagreement between annotators.

Passage	Discussion
Heuristics/Surface-level cues	
In all sincerity, the lesson here is for me to never doubt Fizzy, Connor says, and the audience Awwwwws. "But listen," Lanelle says. "The two of you really had an amazing connection on-screen." Unease thrums beneath my skin. I don't want her to put Connor on the spot like this. "A corpse would have chemistry with this man, Lanelle. Be serious. "The Connor fangirls in the audience scream." No, no, this is something special.	LLM predicts romantic interest between Connor and Fizzy may be due to an incorrect understanding about on-screen vs real connection.
Connor was trying to talk it out with you, Jess says over the steaming top of her mug. I don't need reminding. Every regrettable, overreactive moment of my meltdown is imprinted in my brain like a bad, drunken tattoo. . . . "I know he was. And I know this all happened like eight years ago, and he was upset, and he's older and wiser, but the fact that he decided to not just end his marriage but explode it..." "Fizzy, we are all dumb when we're young. . . .	LLM predicts Fizzy and Connor as ex-spouses potentially due to surface-level cues and improper understanding of who is speaking to/about whom.
Nothing can console him. "My heart is breaking." Why are you telling me this? "Because, as you suggested, I've taken on the role of the Goddess of Women and Children. Do you know what that means?" I shake my head. "It means that everyone who once feared me now loves me. Even Shin, my greatest enemy, loves me. He knows me now as a goddess of motherhood and children. He knows me as a goddess who is loving and kind and giving. Tell me, Mina, how could I be cruel to someone who loves me?" "I do n't know. Can you?" "It's... strange. When I was feared, I hated everything and everyone."	LLM predicts that Mina and Shin are enemies due to misinterpretation of 'me' as 'Mina' instead of the Goddess which shows that it ignores the context and resorts to surface-level cues.
Incorrect (assumed) pronoun resolution	
It's a barely restrained Uh, okay, buddy. It's a laugh held in. Connor's smile remains, but it doesn't look totally natural anymore. "Do you read her books? " Ashley shakes her head. " Oh, I don't read books with just romance in them; I need there to be some plot, too. "Fizzy goes quietly stony." There's plenty of plot. And Fizzy's are the gold standard. "I stare up at him with fondness. This liar, still pretending he's read my books.	LLM assumes "I" to refer to Fizzy resulting in romantic interest prediction between Fizzy and Connor.
So, he says, and smiles shyly over at me in a way that acknowledges how heavy things just got, how there is something hot and tangible in the air between us but maybe if we talk over it, it will dissipate. "You ready for tomorrow?" Inhaling sharply, I sit up straighter. Right. Get yourself together, Fizzy. "I am. I hope I can sleep tonight. I really don't want to show up all puffy and shadowed tomorrow." "I was going to say," he says, smiling, "you've appeared very calm for someone who's about to be on television."	LLM predicts *romantic interest* between Fizzy and Connor even though Connor is not mentioned in the passage.
Sensitivity to irrelevant changes	
"Where is she?" Mom frowned. "Where did you sleep?" Jana rummaged through her bag to get clothes. "Imani's with Anil. They're fine. Everything is fine. I'm just going to take a shower." "But where did you sleep?" Mom asked again. Jana did not want to answer the question. She did not want to say she slept with Anil Malek's arm around her. Or that he wasn't wearing a shirt. Or that they weren't sleeping at all early in the morning and were instead watching the most beautiful sunrise Jana had ever seen and maybe thinking about kissing.	Evolution prediction between Jana and Anil changes from *positive* to *cannot be determined* when *she* is substituted with *Imani*.

Table 9: Examples where LLM's predictions are incorrect due to potential reliance on surface-level heuristics, the tendency to resolve pronouns to the character in question in the absence of an explicit mention, and sensitivity to irrelevant changes in the context.