EvolvTrip: Enhancing Literary Character Understanding with Temporal Theory-of-Mind Graphs

Anonymous ACL submission

Abstract

A compelling portrayal of characters is essential to the success of narrative writing. For readers, appreciating a character's traits requires the ability to infer their evolving beliefs, desires, and intentions over the course of a complex storyline, a cognitive skill known as Theory-of-Mind (ToM). Performing ToM reasoning in prolonged narratives requires readers to integrate historical context with current narrative information, a task at which humans excel but Large Language Models (LLMs) often struggle. To systematically evaluate LLMs' ToM reasoning capability in long narratives, we construct LitCharToM, a benchmark of character-centric questions across four ToM dimensions from classic literature. Further, we introduce EvolvTrip, a perspective-aware temporal knowledge graph that tracks psychological development throughout narratives. Our experiments demonstrate that EvolvTrip consistently enhances performance of LLMs across varying scales, even in challenging extended-context scenarios. EvolvTrip proves to be particularly valuable for smaller models, partially bridging the performance gap with larger LLMs and showing great compatibility with lengthy narratives. Our findings highlight the importance of explicit representation of temporal character mental states in narrative comprehension and offer a foundation for more sophisticated character understanding.

1 Introduction

Theory of Mind (ToM), the capability to infer others' mental states such as beliefs, desires, and intentions, is substantial for narrative comprehension (Premack and Woodruff, 1978; Apperly, 2010), where understanding charaters' motivations and predicting their behaviors across extended storylines demands readers to construct rich mental models of each character. Specifically, ToM reasoning over prolonged narratives requires comprehensive

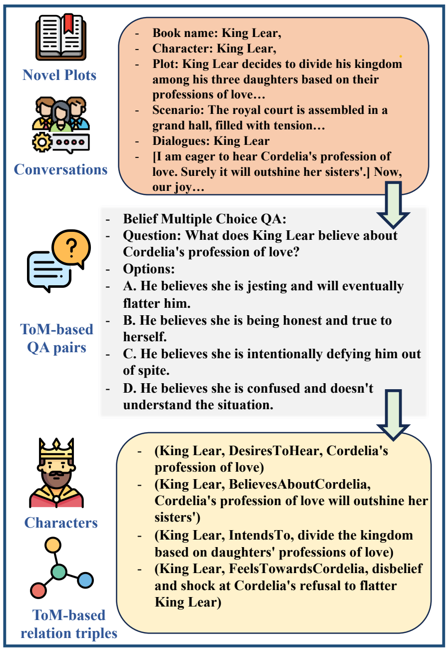

Figure 1: Our ToM-based character understanding pipeline, showing how novel plots and character conversations are transformed into multiple-choice questions and structured relation triples that represent character mental states across belief, desire, intention, and emotion dimensions.

contextualization of accumulated knowledge about characters' backgrounds, personalities, and past experiences with their current circumstances (Davis, 1983; Harwood and Farrar, 2006; Apperly, 2010). When engaging with narratives, humans constantly construct and update models of characters' mental states throughout the storyline, allowing for tracking psychological development and drawing connections between past experiences and present behaviors (Schneider, 2001). Such a temporal and evolutionary dimension of understanding, which is crucial for deep character comprehension, remains underexplored in computational approaches. Despite the increasing sophistication of Large Language Models (LLMs), research reveals significant limitations in their ToM reasoning capabilities, particularly in complex narrative contexts (Nematzadeh et al., 2018b; Gandhi et al., 2023; Tracey et al., 2022; Ullman, 2023; Zhou et al., 2025).

Perspective-taking, which involves inferring what different characters perceive and know based on their unique vantage points, constitutes a critical aspect of human ToM reasoning (Davis, 1983; Harwood and Farrar, 2006). For readers of novels, perspective-taking is enriched by accumulated knowledge of characters' backgrounds and past experiences. However, existing computational approaches to ToM reasoning often neglect this crucial dimension, instead focusing on isolated scenarios without sufficient global context (Wilf et al., 2023; Huang et al., 2024; Hou et al., 2024; Jung et al., 2024; Zhou et al., 2025). Prior ToM benchmarks like CharToM (Zhou et al., 2025) evaluate understanding through brief vignettes with limited character history.

In light of the need for a benchmark that examines LLMs' long-context ToM reasoning capabilities, we construct LitCharToM. LitCharToM is built upon classic literary narratives with characters that possess rich experiences developed over time through multiple interactions and evolving circumstances. This temporal dimension allows us to evaluate models' ability to keep track of characters' psychological evolutions, an essential capability for human-like narrative comprehension.

To enhance LLMs' ToM reasoning capabilities in long narratives, we propose EvolvTrip a novel framework for understanding fictional characters via temporal-aware structured mental state representation. While previous works such as Percept-ToM and EnigmaToM (Jung et al., 2024; Xu et al., 2025) focus on visual perception, EvolvTrip models complex mental states informed by characters' backgrounds, histories, and accumulated experiences. By encoding these perspective-aware mental states as structured triples within a temporal knowledge graph, EvolvTrip enable LLMs to reason about character psychology with contextual richness more closely resembling human ToM processes during narrative comprehension. Empirical results show that EvolvTrip brings significant performance improvements in long-context ToM reasoning to a range of LLMs. EvolvTrip is particularly effective in modeling ToM in extended-context scenarios with corss-plot narrative contents. Further, EvolvTrip is also effective when used with smaller LLMs, partially bridging the performance gap with larger architectures and demonstrating enhanced resilience when processing longer narratives.

Our contributions can be summarised as follows:

- We construct LitCharToM, a character-centric benchmark for evaluating ToM reasoning in literary contexts using classic novels. LitCharToM provides rich scenarios with complex social dynamics and long-term narrative dependencies, enabling comprehensive assessment of contextual understanding.

- We introduce a perspective-aware temporal knowledge graph with entity-guided character linking. Our knowledge graph represents characters' mental states as structured triples tagged with temporal markers and connects character instances across narrative segments.

- We propose EvolvTrip, a neuro-symbolic approach for enhancing ToM reasoning. EvolvTripincorporates structured representation of characters' evolving mental states, which significantly improves LLMs' performance on character-centric ToM reasoning that require deep contextual understanding.

2 Related Work

2.1 Theory of Mind Evaluation in LLMs

Numerous benchmarks have been developed to evaluate ToM capabilities in LLMs by simulating psychological and cognitive experimental designs. Early benchmarks like ToMi (Nematzadeh et al., 2018a) focused on evaluating models' ability to reason about basic beliefs. This foundation was extended by SocialIQA (Sap et al., 2019b), which specifically tests social and emotional intelligence. More advanced ToM reasoning has been explored in Hi-ToM (Wu et al., 2023), which assesses higher-order recursive reasoning about others' beliefs. Recent benchmarks have diversified the evaluation contexts, with FANToM (Kim et al., 2023) stress-testing ToM within conversational settings and OpenToM (Xu et al., 2024) incorporating explicit personality traits and preferences. Comprehensive evaluation platforms like ToMBench (Chen et al., 2024) encompass multiple tasks that target 31 distinct social cognitive abilities. Despite their wide coverage, these benchmarks share common limitations. Most rely heavily on pre-determined rules and templates for scenario generation (Nematzadeh et al., 2018a; Le et al., 2019), which can

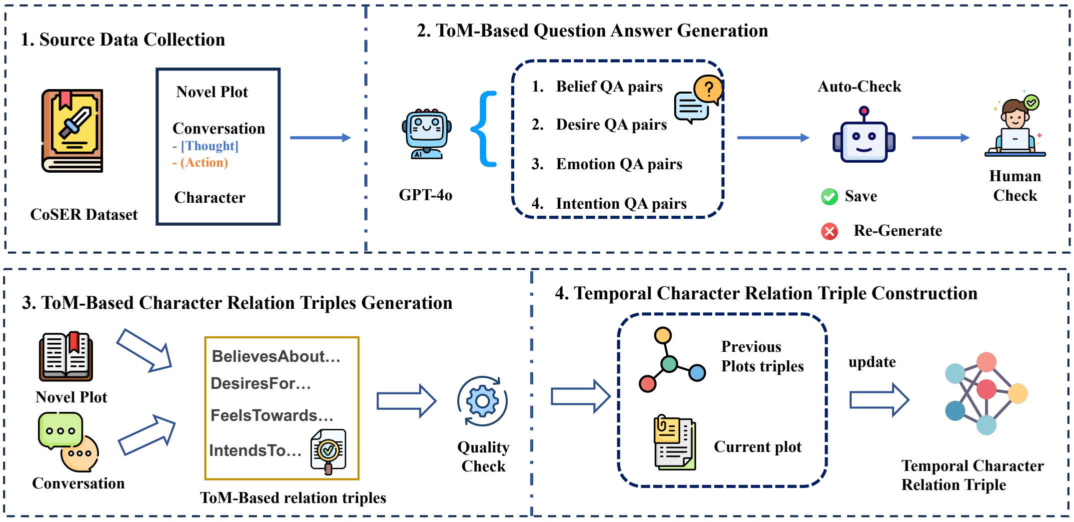

Figure 2: Our ToM-based character understanding pipeline: (1) Source data collection from CoSER Dataset including novel plots and character conversations with [Thought] and (Action) annotations, (2) GPT-4o generation of belief, desire, emotion, and intention QA pairs with two-stage verification, (3) Extraction of BelievesAbout, DesiresFor, FeelsTowards, and IntendsTo relation triples, and (4) Temporal knowledge graph construction by integrating previous and current plot information.

introduce predictable patterns and spurious correlations, potentially leading to the Clever Hans phenomenon (Lapuschkin et al., 2019). Moreover, they typically feature brief, isolated scenarios that fail to capture the complexity of social relationships and interactions that characterize real-world ToM reasoning, overlooking the importance of comprehensive contextual understanding that spans extended narrative timeframes.

Character Understanding in Narrative Comprehension There has been consistent efforts in character-centric narrative understanding, with works like NarrativeQA (Kočiský et al., 2018), LitBank (Bamman et al., 2019; Sims et al., 2019; Bamman et al., 2020), LiSCU (Brahman et al., 2021), and PeQA (Xu et al., 2022) developing question-answering frameworks for longer narrative contexts. These approaches primarily evaluate surface-level comprehension rather than deeper understanding of characters' mental states and psychological development. The psychology literature consistently shows that human readers construct rich mental models of fictional characters' beliefs and intentions (Apperly, 2010), tracking these mental states across extended narratives. This cognitive process relies heavily on accumulated knowledge of characters' backgrounds, histories, and evolving psychological states—aspects that most computational approaches have not adequately modeled.

Knowledge Representation for ToM Reasoning Knowledge bases for representing mental states and social reasoning have evolved from general-purpose semantic networks like ConceptNet (Liu and Singh, 2004) to more specialized representations. Event2Mind (Rashkin et al., 2018) introduced event-based knowledge graphs that capture characters' intentions and reactions, while ATOMIC (Sap et al., 2019a) models if-then relationships for simple social events. Recent approaches include entity state tracking in procedural contexts (Tandon et al., 2020; Zhang et al., 2023), though these have not been specifically applied to character understanding in extended narratives. In the mean time, Neural knowledge bases like COMET is developed (Bosselut et al., 2019), which generate commonsense inferences about social situations, but lack the temporal depth needed for character tracking across narrative arcs.

3 Dynamic Character Understanding through Evolving Mental State Triplets

We introduce the construction of the LitCharToM benchmark and the design of EvolvTrip framework for evaluating Theory-of-Mind comprehension in literary narratives. EvolvTrip (Evolving Triplets) is a structured knowledge representation approach that captures the dynamic evolution of character mental states across narrative arcs. Following the pipeline illustrated in Figure 2, our construction methodology encompasses four integrated phases: (1) source data collection, (2) ToM-based question generation, (3) character relation triple extraction, and (4) temporal knowledge graph construction.

3.1 LitCharToM: Source Data Collection

LitCharToM builds upon the CoSER dataset[1] (Wang et al., 2025), which comprises 81 literary works from project Gutenberg. CoSER provides rich character-centric data including plot summaries, character profiles, and multi-dimensional dialogues. We further selected 20 books from CoSER that exhibit sophisticated character development, complex interpersonal dynamics, and narrative depth spanning multiple scenes. See Appendix A for detailed statistics of LitCharToM.

We base our LitCharToM on CoSER dataset because of its multi-dimensional representation of character dialogue, which includes *verbal speech* (direct communications), *actions* (physical behaviors denoted by parentheses), and *thoughts* (internal cognitive processes denoted by brackets). This tripartite structure offers particular value for ToM analysis, as each dimension maps differently to mental state categories. *Actions* reveal intentions and emotions (e.g., nods firmly suggests deliberate agreement). *Thoughts* provide rich access to all four ToM dimensions, with strongest mapping to emotions (e.g., [I'm terrified]), followed by desires (e.g., [I wish I could leave]), intentions (e.g., [I'll confront him tomorrow]), and beliefs (e.g., [He's lying to everyone]). This structured representation enables EvolvTrip to extract both explicit and implicit mental states from complementary sources, where thoughts reveal deeper affective and cognitive layers, and actions reflect behavioral manifestations of internal states.

3.2 LitCharToM: ToM-Based Question Generation

For each character participating in each plot's dialogues, we systematically generate ToM questions across four dimensions: belief, emotion, intention, and desire. We employ GPT-4o (OpenAI, 2024) to construct multiple-choice questions requiring reasoning about characters' mental states.

For each ToM dimension, GPT-4o examines multiple sources of information: the current plot content, conversation scenario, character dialogues (including the thoughts of current character), and summaries of previous plot segments. This comprehensive context allows the model to identify salient mental states across narrative progression, formulating complex questions with four answer options: one correct answer grounded in the character's depicted psychology and three plausible distractors representing common misinterpretations. To ensure accuracy, we implement a two-stage verification process: initially, GPT-4o verifies all generated questions for logical consistency, clarity, and the presence of a single unambiguously correct answer. Subsequently, human annotators assess accuracy, difficulty level, and appropriateness. Notably, over 90% of the entries are valid at the first generation attempt[2], demonstrating the effectiveness of our generation methodology. Questions identified as problematic during either verification stage undergo refinement or complete regeneration, followed by an additional verification process.

3.3 EvolvTrip: Mental State Triple Extraction

To provide a structured representation of characters' mental activities, EvolvTrip extracts character-centric mental state triples following a subject-predicate-object structure. The subject corresponds to the character, the predicate indicates the ToM dimension (e.g., BelievesAbout, FeelsTowards, IntendsTo, DesiresFor), and the object constitutes the content of the mental state.

For each narrative plot, we employ GPT-4o to generate triples by analyzing the multi-dimensional dialogue data through a perspective-taking lens, which distinguishes between information accessible to each character versus information they cannot know. This perspective-aware approach examines character thoughts that directly reveal mental states, character actions that imply underlying mental states, and verbal dialogues containing explicit statements about beliefs, emotions, intentions, or desires. By identifying events observable by a given character and excluding unobservable ones, this approach significantly alleviates the reasoning burden for LLMs, enabling more accurate mental state attribution. Predicates are specified to provide precise context, such as using BelievesAbout to indicate a belief concerning another entity or FeelsTowards to denote an emotion directed at someone. For triple verification, GPT-4o conducts initial assessment of all generated triples for logical con-

[1]We use the Gutenberg branch of the CoSER dataset to ensure copyright compliance. https://huggingface.co/datasets/Neph0s/CoSER-Books-Gutenberg

[2]See Appendix A.2 for detailed statistics on data quality control.

sistency with the narrative context, adherence to the correct triple format, and appropriate perspective constraints (ensuring characters only form mental states about information they could plausibly access). We then randomly select 40% of triples for human expert verification, assessing their accuracy and relevance to the characters' depicted mental states. Triples identified as incorrect during either verification stage are regenerated and re-verified, ensuring high-quality knowledge representation. Detailed dataset quality statistics are provided in Appendix A.2.

3.4 EvolvTrip: Temporal Knowledge Graph Construction

The core innovation of EvolvTrip is capturing the dynamic nature of character psychology throughout narratives. We construct a temporal knowledge graph where nodes represent characters or significant events, edges embody the generated triples with labels specifying the ToM dimension, and temporal tags associate each triple with specific plot numbers. Each triple is tagged with the plot segment in which the mental state appears, enabling systematic tracking of psychological development. We establish inter-plot links between instances of the same character across different segments, facilitating analysis of how characters' mental states evolve in response to narrative developments.

To maintain psychological consistency, we provide GPT-4o the past mental states of each character when generating triples for new plot segments. This approach enables it to build upon established psychological profiles. For similar mental states concerning the same subject, EvolvTrip combines or refines them based on new information. When new information contradicts earlier states, we update the triples to reflect character development, clearly indicating the temporal transition to demonstrate how the character's perspective has evolved throughout the narrative. This temporally linked representation provides a comprehensive view of character psychology that evolves organically through the narrative, capturing the dynamic nature of beliefs, emotions, intentions, and desires as they transform in response to story events.

4 Experiments

4.1 Setup

We conduct experiments on our multiple-choice Theory-of-Mind benchmark comprising 2,539 questions spanning four dimensions: belief, emotion, intention, and desire. All experiments use a standardized prompt template as detailed in Appendix B. To investigate models' ability to leverage contextual information for ToM comprehension, we vary the context lengths of story plots provided to the models, examining their performance with and without the structured triple representations generated by EvolvTrip . For each question, models are evaluated in two settings: (1) standard prompting with only the narrative context and question, and (2) EvolvTrip -enhanced prompting where relevant mental state triples are included as additional context. This allows us to assess the impact of EvolvTrip's explicit structured knowledge on models' ToM reasoning capabilities.

Evaluated LLMs. We evaluate a diverse set of LLMs as our baselines, including GPT-4o and GPT-4o-mini (OpenAI, 2023), accessed through official APIs. For the open-sourced LLMs, we include DeepSeek-R1 (DeepSeek-AI, 2025), Qwen2.5-72B-Instruct (Yang et al., 2024), Llama3.3-72B-Instruct (Dubey et al., 2024), DS-R1-Dist-Qwen-32B (DeepSeek-R1 distilled into a 32B Qwen architecture) (DeepSeek-AI, 2025), Qwen3-32B (Yang et al., 2025), Qwen2.5-32B-Instruct (Yang et al., 2024), InternLM2.5-20B-Chat(Cai et al., 2024), Qwen3-14B (Yang et al., 2025), Qwen2.5-14B (Yang et al., 2024), DS-R1-Dist-Qwen-14B (DeepSeek-AI, 2025), Qwen3-8B (Yang et al., 2025), Qwen2.5-7B-Instruct (Yang et al., 2024), InternLM3-8B-Instruct (Cai et al., 2024), and InternLM2.5-7B-Chat (Cai et al., 2024). For each model, we test both a standard version and a triple-enhanced version (denoted as "w Triple") that incorporates structured mental state triples into the context. All models are accessed either through official APIs or using weights downloaded from Hugging Face repositories, in compliance with their terms of use.

4.2 Out-of-Distribution Evaluation

To evaluate the generalizability of EvolvTrip to new literary works, we conducted experiments using five books as an out-of-distribution (OOD) test set, comprising 779 questions across the four ToM dimensions. This setup allowed us to assess how well models augmented with EvolvTrip 's structured representations can transfer their ToM reasoning capabilities to entirely new narrative contexts not seen during training or development. For these ex-

Models	Belief Acc.	Desire Acc.	Emotion Acc.	Intention Acc.	Avg Acc.
GPT-4o-mini	66.61	70.06	69.61	71.81	69.52
w Triple	**71.65**	**73.06**	**74.02**	**74.80**	**73.38**
GPT-4	68.35	70.54	72.28	72.28	70.86
w Triple	**71.71**	**73.41**	**75.89**	**75.45**	**74.12**
DeepSeek-R1	68.35	70.91	72.76	71.97	70.74
w Triple	**72.43**	**73.67**	**76.54**	**75.12**	**74.44**
Qwen2.5-72B-Ins.	61.94	63.51	66.05	66.37	64.47
w Triple	**62.58**	**63.04**	**65.73**	**66.21**	**64.39**
Llama3.3-70B-Ins.	61.94	62.73	64.48	65.26	63.60
w Triple	**61.79**	**62.73**	**64.48**	**65.10**	**63.53**
DS-R1-Dist-Qwen-32B	58.65	60.58	62.35	63.17	61.19
w Triple	**62.17**	**63.25**	**65.82**	**66.04**	**64.32**
Qwen3-32B	57.87	60.36	59.91	62.44	60.15
w Triple	**61.39**	**61.89**	**64.28**	**65.25**	**63.21**
Qwen2.5-32B-Ins.	58.82	60.22	61.02	61.97	60.51
w Triple	**60.96**	**63.33**	**65.15**	**66.37**	**63.44**
InternLM2.5-20B-Chat	54.41	56.91	59.61	59.92	57.71
w Triple	**56.78**	**59.16**	**63.37**	**61.04**	**59.53**
Qwen3-14B	56.57	58.54	59.81	60.37	58.04
w Triple	**61.28**	**61.48**	**66.01**	**65.07**	**63.46**
Qwen2.5-14B	57.40	59.44	61.47	60.06	59.64
w Triple	**60.17**	**60.30**	**64.55**	**64.02**	**62.18**
Qwen3-8B	54.88	56.59	59.61	58.50	57.40
w Triple	**59.89**	**61.77**	**64.00**	**63.63**	**62.07**
Qwen2.5-7B-Ins.	57.20	57.07	58.75	58.40	57.87
w Triple	**58.89**	**59.70**	**65.15**	**63.38**	**61.47**
DS-R1-Dist-Qwen-14B	57.15	59.82	60.76	61.25	59.75
w Triple	**61.04**	**61.23**	**65.48**	**64.86**	**63.15**
InternLM3-8B-Instruct	**53.15**	**55.96**	58.03	59.61	56.69
w Triple	51.25	53.53	**61.57**	**62.72**	**57.29**
InternLM2.5-7B-Chat	53.32	55.75	65.18	62.95	59.98
w Triple	**55.32**	**57.75**	**67.18**	**64.95**	**61.98**

Table 1: Multichoice QA accuracy scores of LLMs. The input to LLMs is the current story plots. w / Triple indicates the prompt includes the character's ToM-based relation triples. Best performance of each model is bolded

periments, we selected three representative smaller-scale models: Qwen3-8B, Qwen2.5-7B-Instruct, and InternLM3-8B-Instruct. We evaluated each model in two distinct settings:

Direct Inference. Models were provided with the story plot, conversation scenario description, and question without any fine-tuning. We tested both standard inference (using only narrative content) and EvolvTrip -enhanced inference (including relevant mental state triples in the context).

EvolvTrip-based Fine-Tuning. Models were fine-tuned on training data where the output format first presented the relevant character relation triples followed by the correct answer option. This structured approach was designed to help models learn the explicit connections between narrative information, character mental states, and appropriate answers. The EvolvTrip -based fine-tuning approach offers a significant advantage: it guides models to first extract structured knowledge representations before generating answers, effectively decomposing the complex ToM reasoning process into more manageable steps. By learning to generate structured triples as an intermediate step, models develop a more robust understanding of character psychology that transfers more effectively to new literary contexts. Results from these experiments are presented in Table 3, demonstrating how the EvolvTrip -based approaches impact performance across different model architectures when faced with previously unseen literary works. We provide the training examples in Appendix C.

5 Results and Analysis

5.1 Performance on ToM Reasoning Tasks

The experimental results demonstrate the significant impact of EvolvTrip 's structured mental state triples across various ToM reasoning dimensions. As shown in Table 1, the integration of triple representations consistently enhances model performance, with improvements observed across all model scales and ToM dimensions. With an average prompt length of 2,500 tokens for both standard and EvolvTrip -enhanced inputs, these improvements highlight the value of structured representation rather than simply increasing context length.

The EvolvTrip-enhanced approach yields substantial performance gains for all evaluated mod-

Models	Belief Acc.	Desire Acc.	Emotion Acc.	Intention Acc.	Avg Acc.
GPT-4o-mini	68.66	70.69	72.28	72.59	71.05
w Triple	**71.50**	**73.64**	**75.54**	**75.85**	**74.13**
GPT-4	67.87	71.64	74.17	**75.75**	72.36
w Triple	**70.87**	**72.53**	**75.54**	75.22	**73.54**
DeepSeek-R1	68.76	70.22	72.49	72.43	70.98
w Triple	**71.81**	**73.85**	**75.85**	75.01	**74.13**
Qwen2.5-72B-Ins.	62.50	63.61	66.15	65.99	64.56
w Triple	**63.07**	**64.32**	**67.01**	**66.85**	**65.31**
Llama3.3-70B-Ins.	61.47	63.80	65.77	65.42	64.12
w Triple	**62.76**	**64.27**	**67.17**	**66.69**	**65.22**
DS-R1-Dist-Qwen-32B	66.24	68.35	70.42	71.19	69.05
w Triple	**70.56**	**72.43**	**74.85**	**74.97**	**73.20**
Qwen3-32B	**61.72**	**63.05**	**66.51**	66.37	**64.41**
w Triple	60.91	62.12	66.36	**67.21**	64.15
Qwen2.5-32B	61.81	64.79	66.85	**67.01**	65.12
w Triple	**62.13**	**64.95**	**66.69**	66.85	**65.16**
InternLM2.5-20B-Chat	56.73	58.87	63.84	62.89	60.58
w Triple	**58.30**	**60.42**	**64.44**	**63.32**	**61.62**
Qwen3-14B	52.03	53.41	56.28	56.32	54.51
w Triple	**54.05**	**55.10**	**58.29**	**58.33**	**56.44**
Qwen2.5-14B-Ins.	51.81	52.11	57.17	57.17	54.57
w Triple	**53.81**	**53.80**	**59.17**	**58.69**	**56.37**
Qwen3-8B	49.22	51.76	54.94	55.09	52.75
w Triple	**51.82**	**54.79**	**58.12**	**58.28**	**55.75**
Qwen2.5-7B-Ins.	51.34	52.90	56.54	54.80	53.90
w Triple	**54.02**	**55.74**	**59.54**	**58.28**	**56.90**
DS-R1-Dist-Qwen-14B	53.26	54.89	58.15	58.68	56.25
w Triple	**57.85**	**59.47**	**63.26**	**63.75**	**61.08**
InternLM3-8B-Ins.	50.35	51.95	55.19	55.36	53.21
w Triple	**54.87**	**55.60**	**59.31**	**59.72**	**57.38**
InternLM2.5-7B-Chat	50.35	51.95	55.19	55.36	53.21
w Triple	**54.87**	**55.60**	**59.31**	**59.72**	**57.38**

Table 2: Multichoice QA performances of LLMs in terms of accuracy. The input to LLMs is the current story plots and previous plots' summary. Best performance of each model is bolded.

Models	Belief Acc.	Desire Acc.	Emotion Acc.	Intention Acc.	Avg Acc.
Direct Inference					
Qwen3-8B	51.10	50.58	53.31	53.83	52.21
w Triple	50.77	50.72	55.38	55.90	53.20
Qwen2.5-7B-Ins.	53.85	52.27	57.40	53.33	54.21
w Triple	54.36	52.27	57.44	53.39	54.34
InternLM3-8B-Ins.	50.40	48.64	54.35	52.66	51.51
w Triple	50.81	50.76	54.59	52.97	52.29
Fine-Tuning					
Qwen3-8B	53.22	53.76	54.94	55.09	54.25
w Triple	**59.57**	**57.50**	**58.74**	**56.67**	**58.12**
Qwen2.5-7B-Ins.	55.22	56.29	56.82	56.93	56.32
w Triple	**59.91**	**57.73**	**58.12**	56.93	**58.17**
InternLM3-8B-Ins.	55.40	56.64	57.35	57.26	56.64
w Triple	**58.91**	**58.73**	**58.12**	**58.93**	**58.67**

Table 3: Ablation study results on out-of-distribution testsets across four ToM dimensions. "w Triple" indicates models that use structured triple representation in either inference or training.

els. DeepSeek-R1 shows the most dramatic improvement, increasing from 70.74% to 74.44% when incorporating EvolvTrip triples. Similarly, Qwen3-14B experiences a remarkable improvement of 5.42%, from 58.04% to 63.46%. Even top-performing models like GPT-4o benefit from EvolvTrip integration, improving from 70.86% to 73.36%. These consistent enhancements highlight the fundamental value of EvolvTrip 's structured knowledge representations in ToM reasoning tasks.

The impact of EvolvTrip is particularly pronounced for emotion recognition, where models show the largest accuracy gains. InternLM2.5-7B-Chat improves by 2.00% in emotion accuracy, from 65.18% to 67.18%, while Qwen3-14B sees a remarkable improvement of 6.20%, from 59.81% to 66.01%. This suggests that EvolvTrip 's explicit structured representations effectively bridge the gap between textual cues and the abstract emotional states they signify. Notably, EvolvTrip integration partially mitigates the performance gap between smaller and larger models. While Qwen3-32B outperforms Qwen3-8B by 2.75% in standard settings, this gap narrows when both incorporate EvolvTrip triples. This demonstrates how EvolvTrip 's structured knowledge representations can enhance the reasoning capabilities of smaller models, making sophisticated ToM reasoning more accessible. EvolvTrip integration also helps balance performance across different ToM dimensions. Without triples, models typically perform best on Intention and worst on Belief, with considerable performance disparities. EvolvTrip integration narrows these gaps, providing more consistent reasoning capabilities across all mental state dimensions. For instance, DeepSeek-R1's performance spread between its strongest and weakest dimensions decreases from 4.41% to 4.11% with EvolvTrip enhancement.

5.2 Performance with Extended Context

Table 2 presents model performance when the input is expanded to include both current story plots and summaries of previous plots, increasing the average prompt length to approximately 4,500 tokens. This extended context scenario reveals important insights about model behavior with longer narratives and the continued effectiveness of EvolvTrip integration under more challenging conditions. The addition of previous plot summaries creates a more challenging reasoning environment for all models, with notable performance decreases compared to the current-plot-only scenario in Table 1. For example, Qwen3-14B's accuracy drops substantially from 58.04% to 54.51%, and Qwen3-8B declines from 57.40% to 52.75%. This performance degradation reflects the well-known challenge LLMs face with longer contexts, where relevant information must be identified within a larger text span. The integration of EvolvTrip 's structured mental state triples provides substantial benefits in this more challenging extended context scenario. DS-R1-Dist-Qwen-14B shows a dramatic improvement from 56.25% to 61.08%, while InternLM3-8B-Instruct improves from 53.21% to 57.38%. This demonstrates the robust utility of EvolvTrip 's structured representations in guiding model attention toward relevant character information across longer narrative spans. The benefits of EvolvTrip integration are particularly evident for smaller models, which typically struggle more with extended contexts. Models like Qwen2.5-7B-Instruct show substantial improvements with triples, suggesting that EvolvTrip 's explicit structured knowledge helps these models overcome their inherent limitations in handling longer texts. Performance patterns across ToM dimensions remain consistent with the current-plot-only scenario, with Emotion and Intention dimensions yielding higher accuracy than Belief and Desire dimensions. EvolvTrip integration helps narrow these dimensional performance gaps, providing more balanced reasoning capabilities.

5.3 Ablation Study

To assess the generalizability of EvolvTrip , we conducted an ablation study using five books as out-of-distribution test cases. These books were not part of the training data, allowing us to evaluate how well models transfer ToM reasoning capabilities to entirely new literary contexts. As shown in Table 3, we compare two inference strategies across three model architectures. In the Direct Inference setting, models show modest performance on ToM reasoning tasks, with EvolvTrip -enhanced inference consistently outperforming standard inference across all dimensions. This confirms that EvolvTrip 's structured triple representation provides effective scaffolding for ToM reasoning even without task-specific training. The Fine-Tuning section demonstrates significantly stronger results, where models were trained on data consisting of questions, EvolvTrip 's structured mental state triples, and answers. This triple-based training approach yields substantial improvements across all models and dimensions. For example, Qwen3-8B improves from 54.25% to 58.12% average accuracy when fine-tuned with EvolvTrip triples, and InternLM3-8B-Instruct shows the most dramatic improvement, reaching 58.67% average accuracy. The consistent performance gains across different architectures highlight the transferability of EvolvTrip to novel literary works. Notably, EvolvTrip fine-tuned models maintain balanced performance across all four ToM dimensions, suggesting that the triple-based representation effectively bridges the gap between different types of mental state reasoning.

6 Conclusion

We present EvolvTrip , a structured knowledge representation framework for enhancing Theory-of-Mind reasoning in narrative comprehension. Our character-centric ToM benchmark and perspective-aware temporal knowledge graph transform implicit character psychology into explicit relation triples that evolve throughout narratives. Experiments demonstrate that EvolvTrip significantly enhances reasoning capabilities across model scales and in extended-context scenarios, particularly helping smaller models bridge performance gaps with larger ones.

Ethical Statement

Our benchmark uses literary works from the public domain Gutenberg Project, ensuring proper attribution and copyright compliance. The selected texts span different historical periods and cultural contexts, providing diverse examples of character psychology. Human annotators participating in the verification process were fairly compensated according to standard rates and fully informed about the task nature. We implemented a two-stage verification process to mitigate individual biases in interpretation. We recognise that computational approaches to character understanding inevitably encode particular cultural perspectives or interpretive biases. Literary interpretation varies across cultural traditions, and our framework may reflect Western conceptions of psychology more prominently. While our research aims to advance fundamental capabilities in narrative comprehension, we acknowledge the broader implications for artificial systems that can model human mental states, emphasizing the importance of developing such technologies within frameworks that prioritize transparency and responsible use.

Limitations

Our approach presents several limitations. First, reliance on GPT-4o for triple extraction introduces potential biases in character psychological profiles, as the model may favor certain interpretations over others or miss subtle contextual cues present in the original text. Second, our focus on four ToM dimensions (belief, emotion, intention, desire) doesn't capture other important aspects such as recursive beliefs (beliefs about others' beliefs), counterfactual reasoning, or epistemic states like uncertainty. Third, the structured triple format necessarily simplifies the complex, ambiguous nature of literary character psychology—for instance, a character's conflicted emotions or unconscious motivations may not fit neatly into subject-predicate-object structures. Finally, our multiple-choice evaluation, while allowing for systematic assessment, restricts measurement to recognition rather than testing deeper generative understanding of character psychology.

References

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. 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 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15959–15983. Association for Computational Linguistics.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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*.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Michelle D Harwood and M Jeffrey Farrar. 2006. Conflicting emotions: The connection between affective perspective taking and theory of mind. *British Journal of Developmental Psychology*, 24(2):401–418.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. *arXiv preprint arXiv:2407.01455*.

X Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony Cohn, and Michael Wooldridge. 2024. A notion of complexity for theory of mind via discrete world models. *arXiv preprint arXiv:2406.11911*.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19794–19809, Miami, Florida, USA. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14397–14413. Association for Computational Linguistics.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *CoRR*, abs/1902.10178.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5871–5876. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018a. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2392–2400. Association for Computational Linguistics.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018b. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-02-09, 2024-02-11, 2024-02-12.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

Ralf Schneider. 2001. Toward a cognitive theory of literary character: The dynamics of mental-model construction. *Style*, 35(4):607–639.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi Mishra, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. *arXiv preprint arXiv:2011.08092*.

Jennifer Tracey, Owen Rambow, Claire Cardie, Adam Dalton, Hoa Trang Dang, Mona T. Diab, Bonnie J. Dorr, Louise Guthrie, Magdalena Markowska, Smaranda Muresan, Vinodkumar Prabhakaran, Samira Shaikh, and Tomek Strzalkowski. 2022. Best: The belief and sentiment corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation*

Conference, LREC 2022, Marseille, France, 20-25 June 2022, pages 2460–2467. European Language Resources Association.

Tomer David Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *ArXiv*, abs/2302.08399.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, et al. 2025. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10691–10706. Association for Computational Linguistics.

Hainiu Xu, Siya Qi, Jiazheng Li, Yuxiang Zhou, Jinhua Du, Caroline Catmur, and Yulan He. 2025. Enigmatom: Improve llms' theory-of-mind reasoning capabilities with neural knowledge base of entity states. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8593–8623. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Li Zhang, Hainiu Xu, Abhinav Kommula, Chris Callison-Burch, and Niket Tandon. 2023. Openpi2.0: An improved dataset for entity tracking in texts. *arXiv preprint arXiv:2305.14603*.

Chulun Zhou, Qiujing Wang, Mo Yu, Xiaoqian Yue, Rui Lu, Jiangnan Li, Yifan Zhou, Shunchi Zhang, Jie Zhou, and Wai Lam. 2025. The essence of contextual understanding in theory of mind: A study on question answering with story characters. *arXiv preprint arXiv:2501.01705*.

A Dataset Statistical

A.1 Book Selection and Characteristics

We selected 20 books from the CoSER dataset for the construction of our LitCharToM benchmark. These books from the Gutenberg Project are publicly accessible and span different historical periods, literary styles, and genres. Table A1 lists the chosen books along with their plot counts, conversation counts, and average character numbers. Our benchmark features a diverse collection of 258 plots containing 599 conversations across these works. Notably, these books encompass a wide range of characters crafted by different authors with varying literary traditions. These characters possess distinct personalities, motivations, and backgrounds, representing diverse psychological profiles from ambitious royalty to contemplative philosophers. This diversity helps mitigate potential biases related to literary style, historical period, and cultural perspective while ensuring comprehensive coverage of different ToM reasoning challenges across narrative contexts. The statistics for books we selected in this paper are shown in Table A1 and Table A2. **Detailed statistics of LitCharToMis shown in Table A3**

A.2 Dataset Quality Control

To ensure data quality, we conduct a rigorous two-stage verification process for both questions and character relation triples. For the ToM-based questions, GPT-4o first verifies all generated questions for logical consistency, clarity, and the presence of a single unambiguously correct answer. Subsequently, human annotators assess a substantial portion of the questions for accuracy, difficulty level, and appropriateness, achieving a verification accuracy of 92.47%. For the triple extraction, we employ a similar two-stage approach, with GPT-4o conducting an initial assessment followed by human expert verification of 40% randomly selected triples, resulting in 93.64% accuracy. Questions or triples identified as problematic during either verification stage undergo refinement or complete regeneration, followed by an additional verification cycle. This iterative process ensures the reliability and correctness of our benchmark for evaluating ToM reasoning capabilities in literary contexts.

A.3 LitCharToM Dataset Statistics

Our LitCharToM benchmark comprises a diverse collection of literary content for evaluating ToM reasoning capabilities. The dataset includes 20 books spanning different literary periods and genres, with 2,539 multiple-choice questions focused on character psychology. Each question is accompanied by one correct answer and three plausible distractor options, resulting in a total of 10,156 answer choices (2,539 correct answers and 7,617 distractors).

We evaluate models in two context settings: standard and extended. In the standard setting (current plot only), the average context length is 2,109 tokens, with a median of 2,094 tokens. For the extended setting (including previous plot summaries), the average context length increases substantially to 4,524 tokens, with contexts ranging from 1,259 to 20,366 tokens. This range of context lengths allows us to systematically evaluate how models handle ToM reasoning across different narrative scopes.

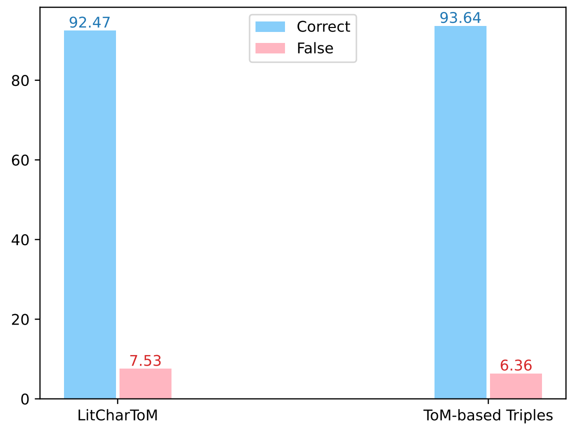

Figure A1: Evaluation of generated data quality for LitCharToM dataset and ToM-based triples. Correct refers to the data verified as accurate by human annotators.

B Prompts

B.1 Prompt for Multiple Choice Question Generation

The prompt for ToM-based multiple choice question generation is shown in Table A5.

B.2 Prompt for Character Relation Triple Generation

The prompt for ToM-based character relation triple generation is shown in Table A6.

Book Name	Plots Num	Conversations Num	Avg Character
King Lear	14	42	3.00
A Study in Scarlet (Sherlock Holmes, #1)	14	41	2.73
The Scarlet Letter	11	37	3.36
The Taming of the Shrew	10	29	2.90
The Merchant of Venice	11	33	3.00
The Tempest	7	23	3.29
Julius Caesar	7	20	2.86
The Call of the Wild	8	22	2.75
A Portrait of the Artist as a Young Man	12	30	2.50
The Wind in the Willows	14	37	2.64
A Little Princess	14	31	2.21
The Importance of Being Earnest	14	36	2.57
Othello	9	26	2.36
Dr Jekyll and Mr Hyde	9	20	2.00
The Hound of the Baskervilles	15	47	2.61
Notes from Underground	19	37	1.85
The Turn of the Screw	20	42	2.10
Jude the Obscure	24	48	2.00
Siddhartha	15	30	2.00
Anthem	11	18	1.64
Total	258	599	2.47

Table A1: Statistics for the 20 books used in the evaluation.

Book Name	Plots Num	Conversations Num	Avg Character
The Hound of the Baskervilles	15	47	2.61
Notes from Underground	19	37	1.85
The Turn of the Screw	20	42	2.10
Jude the Obscure	24	48	2.00
Siddhartha	15	30	2.00
Total	93	204	2.19

Table A2: Statistics for the 5 books used as out-of-distribution test set.

Dataset Characteristics	**Count/Value**
Books	20
Questions	2,539
Correct Answers	2,539
Distractor Answers	7,617

Table A3: Core statistics of the LitCharToM dataset.

Context Length	**Standard Setting**	**Extended Setting**
Average	2,109	4,524
Median	2,094	2,894
Minimum	1,734	1,259
Maximum	2,601	20,366

Table A4: Context length statistics across different evaluation settings.

C Dataset Examples

C.1 OOD Evaluation Results

Table A7 presents detailed model predictions for a representative question from our OOD test set, demonstrating how EvolvTrip 's structured triples influence model reasoning. When comparing models with and without triple information, we observe that triple-enhanced models consistently identify Siddhartha's deeper spiritual intentions more accurately. While InternLM3-8B generates the correct answer even without triples, Qwen3-8B and Qwen2.5-7B-Ins only arrive at the correct answer when provided with explicit triple representations.

This pattern illustrates how EvolvTrip 's structured knowledge helps bridge reasoning gaps, particularly for complex questions requiring nuanced understanding of character motivations across extended narrative contexts.

C.2 Training Set

The training examples for two different experiment setting for OOD evaluation are shown in Table A8 to Table A9.

Prompt for Multiple Choice Question Generation

You are an expert in narrative analysis and character psychology, specializing in the application of Theory of Mind (ToM).

Your task is to generate one multiple choice question for each of the following ToM dimensions — Belief, Emotion, Intention, and Desire — based on the provided story plot, scenario, character dialogues, and previous character relation triples. Each question must probe the psychological state of the Target Character, supported by reasoning grounded in both previously identified mental state triples and the current context.

Definitions of Theory of Mind Dimensions:

<Belief>: What the character believes to be true — this includes both objective facts and their subjective perceptions.

<Emotion>: What the character feels — their affective responses, including joy, anger, fear, etc.

<Intention>: What the character plans or wants to do — immediate or long-term actions driven by goals or motivations.

<Desire>: What the character yearns for or wants to obtain — internal wishes, cravings, or goals (emotional or material).

Input Fields:

Plot summary: Contextual background of the narrative.
Current Scenario: The specific scene or moment in focus.
Dialogues: The words spoken and actions taken by characters in the scene.
Target Character: The character whose mental states are being analyzed.
Previous Character Relation Triples: Previously established mental state triples for the target character.

Output Instructions:
1. For each ToM dimension, select relevant mental state triples.

2. Construct one complex multiple choice question that requires reasoning and inference, not surface recall.

3. Provide four answer options:
- One correct answer, grounded in the character's psychology.
- Three plausible but incorrect distractors, based on common misreadings or partial understanding.

4. Do not repeat the same idea across different options.

Output Format:
```
{
 "Target Character": [
  {"Belief Multiple Choice Question": {
   "Scenario": "xxx", "Reasoning":"xxx", "Question": "xxx",
   "Options": ["A.xxx", "B.xxx", "C.xxx", "D.xxx"],
   "Correct Answer": "x"}},
  {"Emotion Multiple Choice Question": {...}},
  {"Intention Multiple Choice Question": {...}},
  {"Desire Multiple Choice Question": {...}}
 ]
}
```

Table A5: Prompt for Multiple Choice Question Generation.

Prompt for Character Relation Triple Generation

You are an expert in analyzing narrative texts and understanding character psychology through the lens of Theory of Mind. Your task is to extract the triples of beliefs, emotions, intentions, and desires of a specific target character from the provided story plot summary, current scenario, and dialogues. You will output each identified mental state as a subject-predicate-object triple.

Here are the definitions of the Theory of Mind dimensions you should use:

ToM dimensions: <Belief>: Beliefs encompass both objective facts and subjective perceptions concerning the existence or truth of something.

<Emotion>: Emotions are strong feelings deriving from one's circumstances, mood, or relationships with others. And emotions are variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure.

<Intention>: Intentions are blueprints that steer actions, encompassing both future plans and the motivations driving current behaviour.

<Desire>: Desires encompass both physical needs and psychological yearnings. Desires incline people toward action and fulfilling desires is pleasurable. Their fulfillment is normally experienced as pleasurable in contrast to the negative experience of failing to do so.

Analyze the provided Dialogues, the Target Character's explicitly stated Thoughts (if available in square brackets), and their Actions (if available in parentheses) within the context of the Story Plot Summary and Current Scenario.

Identify instances of the Target Character's Beliefs, Emotions, Intentions, and Desires based on the definitions provided above. Output each identified mental state as a triple in the format: (Target Character, Predicate, Object). Predicate should clearly indicate the ToM dimension (e.g., Believes, Feels, Intends, Desires) and can include a brief description of the target of the mental state (e.g., Believes about Cordelia's silence). Object should be the content of the mental state (e.g., Cordelia's silence is a sign of disrespect and rebellion).

This predicate can be further specified to provide more context, for example, using "BelievesAbout" to indicate a belief concerning another entity or event, or "FeelsTowards" to denote an emotion directed at someone or something.

Prioritize information that is directly attributable to the Target Character through their explicitly stated thoughts, actions, or spoken words.

Example

<Plot summary>

In King Lear's palace, Kent and Gloucester discuss the King's preference between Albany and Cornwall. Lear, deciding to divide his kingdom among his daughters, ...

<Current Scenario>

In the opulent grand hall of King Lear's palace, anticipation hangs thick in the air. ...

Dialogues between characters:

Environment: King Lear's grand hall, with courtiers and family gathered, as Lear prepares to speak."

"King Lear: [I must know which daughter loves me most.] Tell me, my daughters, which of you shall we say doth love us most?

"Goneril: Sir, I love you more than words can wield the matter; dearer than eyesight, space and liberty. Cordelia: (remains silent)"

"King Lear: [She speaks well.] Of all these bounds, we make thee lady. What says our second daughter, Regan?"

"Regan: I am made of that self metal as my sister, and prize me at her worth.Cordelia: Then poor Cordelia! And yet not so; since I am sure my love's more ponderous than my tongue. ...

Target Character: King Lear

Output:

{{ "Target Character": [(King Lear, DesiresToKnow, which daughter loves King Lear most), (King Lear, IntendsTo, divide the kingdom based on his daughters' declarations of love), (King Lear, BelievesAboutCordelia, Cordelia's silence is a sign of defiance and disrespect), (King Lear, FeelsTowardsCordelia, wounded and betrayed by Cordelia's refusal to flatter King Lear), (King Lear, BelievesAboutGoneril, Goneril speaks well and expresses her love convincingly), (King Lear, FeelsTowardsCordelia, disappointed and shocked by Cordelia's honesty)] }}

Based on the provided Story Plot Summary, Current Scenario, and Dialogues, identify all relevant beliefs, emotions, intentions, and desires of the Target Character Do not use pronoun in Object, use the name of Target character instead of his/her/them Output each identified mental state as a triple in the format like (Target Character, Predicate, Object) in the following format.

Input

<Plot summary>

<Current Scenario>

Dialogues between characters:

Target Character:

Previous Character Triples:

When analyzing the current scenario, consider the character's previously identified mental states triples from earlier plots. Your task is to:

1. Integrate previous triples with your current analysis 2. For similar predicates (e.g., multiple beliefs about the same subject), combine or refine them based on new information 3. For conflicting predicates, update with the current information to reflect character development 4. Maintain consistency in the character's psychological profile while acknowledging changes in their mental states

Use double quotes for all keys and values in the JSON. Do NOT include any explanation, markdown formatting, or additional comments Only return the JSON object. You MUST return the result strictly in JSON format:

Output: {{ "Target Character": ["(Target Character, Predicate, Object)", "(Target Character, Predicate, Object)", "(Target Character, Predicate, Object)", "(Target Character, Predicate, Object)"] }}

Table A6: Prompt for Character Relation Triple Generation.

OOD Evaluation Input and Gold Triples

You are an expert in narrative analysis and character psychology, specializing in Theory of Mind (ToM). Your task is to analyze the mental states of characters in literary works.

For the character "Siddhartha" in the book "Siddhartha", analyze their mental state based on the following context:

STORY PLOT:
Siddhartha, a Brahmin's son, grows up with his friend Govinda. He excels in spiritual practices and is loved by all. However, he becomes dissatisfied with traditional teachings and seeks a deeper understanding of the self and the universe.

SCENARIO:
In the serene landscape of ancient India, beneath the sprawling branches of a majestic banyan tree, Siddhartha and Govinda sit in a tranquil meditation space. The early morning light filters through the leaves, casting soft shadows on the ground. The surrounding area is peaceful, with distant sounds of chirping birds and a gentle breeze rustling the leaves. The setting represents a pivotal moment in their spiritual journey, where traditional teachings are about to be challenged by Siddhartha's profound inner questioning.

QUESTION:
What is Siddhartha's intention regarding his spiritual journey?
CANDIDATE CHOICES:
A. To continue following traditional teachings without question.
B. To seek a deeper understanding of Atman beyond words and rituals.
C. To convince Govinda to abandon their current path together.
D. To establish himself as a new spiritual leader among the Brahmins.

(Siddhartha, DesiresToExpress, doubts to Govinda without shaking Govinda's faith)
(Siddhartha, Feels, weight of questions growing during meditation)
(Siddhartha, BelievesAboutTeachings, traditional teachings are profound but lacking)
(Siddhartha, Desires, to truly know Atman beyond words and rituals)
(Siddhartha, FeelsTowardsGovinda, touched by Govinda's loyalty but worried for Govinda)
(Siddhartha, IntendsToSeek, further understanding beyond traditional teachings)
(Siddhartha, BelievesAboutPath, true enlightenment might lie beyond current teachings)
(Siddhartha, DesiresToFindTruth, whatever the cost)

What is Siddhartha's intention regarding his spiritual journey?

InternLM3-8B-Ins.	**InternLM3-8B-Ins. (w triples)**	**Qwen3-8B**
Relevant mental state triples: (Siddhartha, IntendsTo, express his inner doubts and desire for deeper understanding to Govinda) (Siddhartha, IntendsTo, seek the ultimate truth beyond words and rituals) Answer:{answer: C}	Relevant mental state triples: (Siddhartha, IntendsTo, explore the nature of the self beyond traditional teachings) Answer:{answer B}	Relevant mental state: (Siddhartha, IntendsTo, seek a deeper understanding of Atman and the universe) (Siddhartha, IntendsTo, set aside dogue's teachings and embrace the path of light together with Govinda) Answer:{answer: C}
Qwen3-8B (w triples)	**Qwen2.5-7B-Ins.**	**Qwen2.5-7B-Ins. (w triples)**
Relevant mental state: (Siddhartha, IntendsTo, understand the true self beyond traditional teachings) (Siddhartha, IntendsTo, approach the spiritual journey with an open mind) Answer:{answer: B}	Relevant mental state triples: (Siddhartha, IntendsTo, question the nature of Atman and the purpose of life) (Siddhartha, IntendsTo, embark on a new path of self-discovery and spiritual exploration) Answer:{answer: A}	Relevant mental state triples: (Siddhartha, IntendsTo, question the teachings of the oldest Brahmin) (Siddhartha, IntendsTo, pursue the understanding of Atman beyond words and rituals) Answer:{answer: B}

Table A7: Model predictions for book Siddhartha's intention question of OOD evaluation.

Training example for OOD evaluation

Input:

You are an expert in narrative analysis and character psychology, specializing in Theory of Mind (ToM). Your task is to analyze the mental states of characters in literary works.

For the character King Lear in the book King Lear, analyze their mental state based on the following context:

STORY PLOT:
King Lear decides to divide his kingdom among his three daughters based on their professions of love. Cordelia, refusing to flatter, is disinherited. Kent is banished for defending her. Lear gives his power to Goneril and Regan, who soon begin to undermine his authority. The Fool criticizes Lear's actions, and Lear starts to realize his mistake. Goneril demands Lear reduce his retinue, leading to a heated confrontation. Lear decides to leave for Regan's castle.

SCENARIO:
The royal court is assembled in a grand hall, filled with tension and anticipation. Ornate tapestries and gilded decorations surround the throne where King Lear sits, preparing to divide his kingdom. Goneril and Regan have already delivered their exaggerated declarations of love, receiving large portions of the kingdom. Now, all eyes are on Cordelia, the youngest daughter, as she prepares to speak. The atmosphere is charged with expectation, silence hanging heavy as courtiers watch the pivotal moment that will determine the future of the kingdom.

QUESTION:
What does King Lear believe about Cordelia's profession of love?

CANDIDATE CHOICES:
A. He believes she is jesting and will eventually flatter him.
B. He believes she is being honest and true to herself.
C. He believes she is intentionally defying him out of spite.
D. He believes she is confused and doesn't understand the situation.

First, identify the relevant mental state triples (beliefs, emotions, intentions, or desires) that explain King Lear's psychology in this scenario.
Then, based on these mental states, select the most appropriate answer from the choices above.

Format your response as:
1. List the relevant mental state triples
2. Provide your answer as a JSON object: {answer: X} where X is the letter (A, B, C, or D) of the correct choice.

Output:

"Relevant mental state triples:
(Siddhartha, IntendsTo, express his inner doubts and desire for deeper understanding to Govinda)
(Siddhartha, IntendsTo, seek the ultimate truth beyond words and rituals)
Answer:
{answer: B}

Table A8: Example of training data with triples.

Training example for OOD evaluation
Input: You are an expert in narrative analysis and character psychology, specializing in Theory of Mind (ToM). Your task is to analyze the mental states of characters in literary works. For the character King Lear in the book King Lear, analyze their mental state based on the following context: STORY PLOT: King Lear decides to divide his kingdom among his three daughters based on their professions of love. Cordelia, refusing to flatter, is disinherited. Kent is banished for defending her. Lear gives his power to Goneril and Regan, who soon begin to undermine his authority. The Fool criticizes Lear's actions, and Lear starts to realize his mistake. Goneril demands Lear reduce his retinue, leading to a heated confrontation. Lear decides to leave for Regan's castle. SCENARIO: The royal court is assembled in a grand hall, filled with tension and anticipation. Ornate tapestries and gilded decorations surround the throne where King Lear sits, preparing to divide his kingdom. Goneril and Regan have already delivered their exaggerated declarations of love, receiving large portions of the kingdom. Now, all eyes are on Cordelia, the youngest daughter, as she prepares to speak. The atmosphere is charged with expectation, silence hanging heavy as courtiers watch the pivotal moment that will determine the future of the kingdom. QUESTION: What does King Lear believe about Cordelia's profession of love? CANDIDATE CHOICES: A. He believes she is jesting and will eventually flatter him. B. He believes she is being honest and true to herself. C. He believes she is intentionally defying him out of spite. D. He believes she is confused and doesn't understand the situation.   First, identify the relevant mental state triples (beliefs, emotions, intentions, or desires) that explain King Lear's psychology in this scenario. Then, based on these mental states, select the most appropriate answer from the choices above. Format your response as: 1. List the relevant mental state triples 2. Provide your answer as a JSON object: {answer: X} where X is the letter (A, B, C, or D) of the correct choice. ## Output: Answer: {answer: B}

Table A9: Example of training data w/o triples.