

# EnigmaToM: Improve LLMs' Theory-of-Mind Reasoning Capabilities with Neural Knowledge Base of Entity States

Hainiu Xu<sup>♣</sup> Siya Qi<sup>♣</sup> Jiazheng Li<sup>♣</sup> Yuxiang Zhou<sup>♣,♠</sup>  
Jinhua Du<sup>♡</sup> Caroline Catmur<sup>♣</sup> Yulan He<sup>♣,◇</sup>

<sup>♣</sup>King's College London

<sup>♡</sup>Huawei London Research Centre

<sup>◇</sup>The Alan Turing Institute

<sup>♠</sup>Queen Mary University of London

{hainiu.xu, yulan.he}@kcl.ac.uk

## Abstract

Theory-of-Mind (ToM), the ability to infer others' perceptions and mental states, is fundamental to human interaction but remains challenging for Large Language Models (LLMs). While existing ToM reasoning methods show promise with reasoning via perceptual perspective-taking, they often rely excessively on off-the-shelf LLMs, reducing their efficiency and limiting their applicability to high-order ToM reasoning. To address these issues, we present EnigmaToM, a novel neuro-symbolic framework that enhances ToM reasoning by integrating a Neural Knowledge Base of entity states (Enigma) for (1) a psychology-inspired *iterative masking* mechanism that facilitates accurate perspective-taking and (2) *knowledge injection* that elicits key entity information. Enigma generates structured knowledge of entity states to build spatial scene graphs for belief tracking across various ToM orders and enrich events with fine-grained entity state details. Experimental results on ToMi, HiToM, and FAN-ToM benchmarks show that EnigmaToM significantly improves ToM reasoning across LLMs of varying sizes, particularly excelling in high-order reasoning scenarios<sup>1</sup>.

## 1 Introduction

Theory-of-Mind (ToM), the ability to understand that others have perceptions and mental states different from one's own, is fundamental to effective communication and social interaction (Premack and Woodruff, 1978; Apperly, 2010). ToM reasoning can be first-order, involving the understanding of another's mental state, or higher-order, requiring recursive thinking about others' beliefs. Higher-order ToM reasoning is particularly vital in real-world contexts such as negotiation (De Weerd et al., 2017). As Large Language Models (LLMs) become increasingly sophisticated in imitating human

<sup>1</sup>The neural knowledge base Enigma can be downloaded via <https://huggingface.co/SeacowX/Enigma>. Code and data are available at <https://github.com/seacowx/EnigmaToM>.

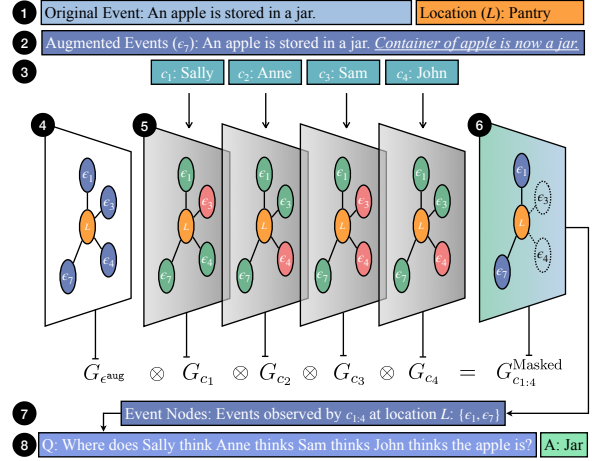


Figure 1: Example use-case of EnigmaToM framework in fourth-order ToM reasoning. An *event* (1) is enriched by adding information about entity-of-interests (*italic text* in (2)) derived from Enigma. Characters (3) are extracted using an off-the-shelf NER model. Spatial scene graphs (4) and (5) are constructed for perspective-taking through a masking mechanism (5 → 6). Event nodes are retrieved to construct character-centric event sequence (7), which is used for the final QA (8).

interactions, a plethora of studies have investigated LLMs' abilities to conduct ToM reasoning. While early studies show that LLMs exhibit traces of ToM capabilities (Bubeck et al., 2023; Kosinski, 2023), follow-up works impugn the robustness of such capabilities by showing that LLMs' ToM reasoning is often superficial (Sap et al., 2022; Ullman, 2023; Shapira et al., 2024).

A vital prerequisite for human ToM reasoning is *perceptual perspective-taking* (referred to as "perspective-taking" thereafter), which is the process of inferring the perception of other characters (Davis, 1983; Harwood and Farrar, 2006). In the case of ToM reasoning with LLMs, perspective-taking alleviates the reasoning burden of LLMs by identifying events that are observable by a given character and removing unobservable ones.

Centered around perspective-taking, numerous methods have been proposed. SimulatedToM (Wilf

et al., 2024) and Discrete World Models (DWM) (Huang et al., 2024) perform perspective-taking by directly prompting LLMs. While one may appreciate these methods’ simplicity, the quality of perspective-taking is largely dependent on the capability of LLMs. SymbolicToM, TimeToM, and PerceptToM took a neuro-symbolic approach. TimeToM (Hou et al., 2024) and PerceptToM (Jung et al., 2024) utilize temporal and perceptual information of events to derive characters’ perception by extracting common timestamps or perceived characters. However, accurately extracting perceived timestamps or perceivers becomes difficult as the length or complexity of the event trajectory increases. The most relevant work to ours is SymbolicToM, where perspective-taking is conducted by maintaining multiple belief graphs (Sclar et al., 2023). However, SymbolicToM constructs belief graphs using less powerful models including WANLI (Liu et al., 2022) and OpenIE (Stanovsky et al., 2018), limiting its generalizability to ToM tasks that involve complicated events. Further, as noted by Sclar et al. (2023), SymbolicToM lacks efficiency as the depth of ToM reasoning increases (see §3.4 for analysis).

Given the need for accurate and efficient perspective-taking in ToM reasoning, we introduce **Entity-Guided Masking** (EnigmaToM), a neuro-symbolic framework enhancing LLMs’ ToM reasoning (Figure 1). Perspective-taking relies on reasoning about event implications, where information about the states of key entities is crucial (Zhang et al., 2023). EnigmaToM employs a Neural Knowledge Base (Enigma) to generate structured entity-state information (§3.1). This entity-state information supports spatial scene graph construction for perspective-taking (§3.3) and event elicitation through knowledge injection (§3.2). Experiment results show that EnigmaToM improves the ToM reasoning capabilities of a range of LLMs. Furthermore, the iterative masking mechanism, grounded by theories from psychology (Arslan et al., 2017), guarantees the efficacy of EnigmaToM across ToM reasoning of varying orders.

We summarize our contributions as follows:

1. We introduce EnigmaToM, a neuro-symbolic framework for ToM reasoning that leverages a Neural Knowledge Base of Entity States to improve LLMs’ ToM reasoning capabilities.
2. Through the iterative masking mechanism, EnigmaToM conducts effective perspective-taking while greatly reducing the number of

character belief graphs that need to be tracked, thereby improving the efficiency in high-order ToM reasoning.

3. EnigmaToM improves LLMs’ ToM reasoning, especially for higher-order cases. Analysis show that EnigmaToM improves LLMs’ ToM reasoning ability up to the fourth order.

## 2 Related Work

**Knowledge Base of Commonsense Knowledge in Natural Language** Efforts to construct commonsense knowledge bases have a long history. Early work includes CyC, ConceptNet, and DBPedia (Lenat, 1995; Liu and Singh, 2004; Lehmann et al., 2015). Rashkin et al. (2018) introduced Event2Mind, an event-based knowledge graph that captures characters’ intentions and reactions. Subsequently, Sap et al. (2019) introduced ATOMIC, a commonsense knowledge graph that models if-then relationships for simple events. To explore more complex events, Tandon et al. (2020) introduced OpenPI, a dataset for entity state tracking in procedures. OpenPI was extended to OpenPI2.0 by introducing entity saliency scores and entity canonicalization (Zhang et al., 2024). Parallel efforts have developed neural models, including a GRU-based encoder-decoder model for Event2Mind (Rashkin et al., 2018), a decoder-only Transformer called COMET for ConceptNet and ATOMIC (Bosselut et al., 2019), and fine-tuned GPT-2 for OpenPI (Tandon et al., 2020).

**Benchmarking LLMs’ ToM Reasoning Capabilities** Many ToM benchmarks are inspired by the False Beliefs test (Wimmer and Perner, 1983), including event-based benchmarks such as ToMi (Le et al., 2019), HiToM (Wu et al., 2023), BigToM (Gandhi et al., 2024), and OpenToM (Xu et al., 2024), and dialogue-based datasets such as FAN-ToM (Kim et al., 2023). Based on the Smarties Test (Gopnik and Astington, 1988), Adv-CSFB (Shapira et al., 2024) and ToMChallenges (Ma et al., 2023) assess LLMs’ ability to reason about unexpected contents and unexpected transfers. ToMBench (Chen et al., 2024) and EPITOME (Jones et al., 2023) contain a suite of ToM tasks that go beyond False Beliefs and Smarties Test. MMTOM-QA extends ToM evaluation to multimodality (Jin et al., 2024) and InformativeBench evaluates ToM in multi-agent settings (Liu et al., 2024).

**Improving LLMs’ ToM Reasoning Capabilities** Methods for improving LLMs’ ToM reasoning ca-

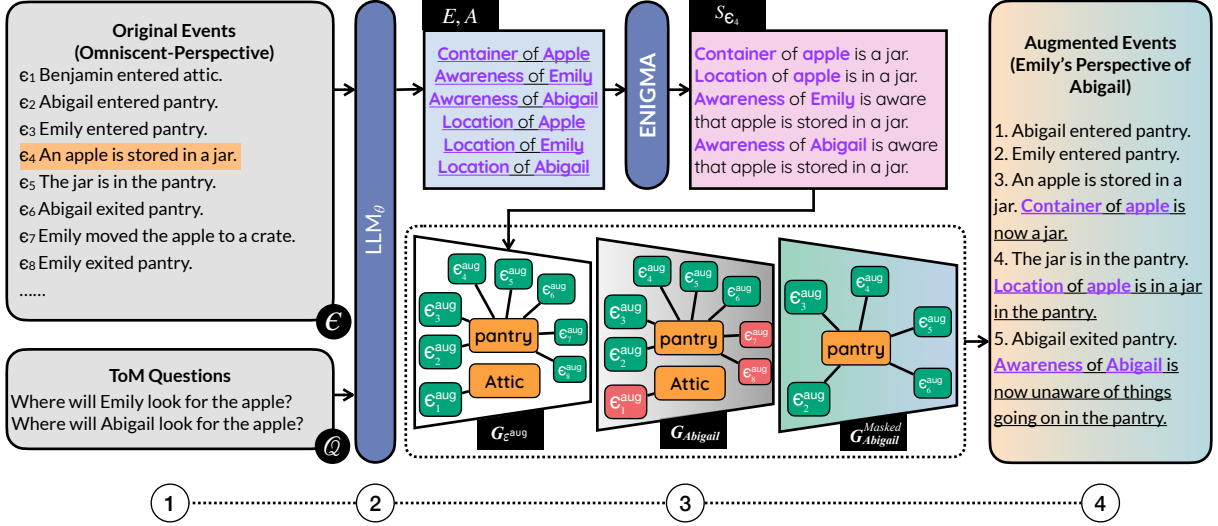


Figure 2: An overview of the EnigmaToM framework. In the graphs shown at bottom of ③, ● nodes denotes observed events while ● nodes denotes unobserved events. See detailed explanations in §3.

pabilities have focused on *perspective-taking*. SymbolicToM conducts perspective-taking via belief graphs (Sclar et al., 2023). SimulatedToM (Wilf et al., 2024) and DWM (Huang et al., 2024) conduct perspective-taking by prompting. DWM additionally prompts LLMs to infer the world state after a group of events. TimeToM utilizes the temporal order of events to conduct perspective-taking (Hou et al., 2024). PerceptToM does perspective-taking by prompting LLMs to infer perceivers of each event (Jung et al., 2024). For multimodal ToM, methods like NIPE and BIP-ALM leverage Bayesian Inverse Planning (Ying et al., 2023; Jin et al., 2024), with environments (e.g. 2D grids or videos) providing strong perspective-taking signals through observable trajectories.

### 3 The EnigmaToM Framework

Before presenting the EnigmaToM framework, we define the general setup of ToM reasoning tasks.

**ToM Task Setup** We focus on the widely studied ToM task of *reasoning about false beliefs* (Wimmer and Perner, 1983), which is typically formulated as QA tasks. Formally, given a context consisting of a sequence of events,  $\mathcal{E} = \{\epsilon_i\}_{i=1}^n$ , which involves multiple characters,  $\mathcal{C} = \{c_j\}_{j=1}^m$ , and a query regarding the belief of a particular character,  $q_c, c \in \mathcal{C}$ , the goal is to derive the most likely belief,  $b_c$ , from all potential beliefs,  $\mathcal{B}_c$ :

$$b_c^* = \arg \max_{b \in \mathcal{B}_c} \mathbb{P}(b | \mathcal{E}, q_c, c \in \mathcal{C}) \quad (1)$$

Further,  $\mathcal{E}$  can be concise events as seen in the ToMi dataset (Le et al., 2019) or utterances as seen

in the FANToM dataset (Kim et al., 2023). Beyond directly querying a character’s beliefs about the environment, one can also probe their beliefs regarding other characters’ perceptions, thereby enabling the assessment of higher-order ToM reasoning.

**The EnigmaToM Framework** Figure 2 provides an overview of our framework, we use circled number (③) to refer to components in the figure. At the core of EnigmaToM is a Neural Knowledge Base (NKB) of Entity States (Enigma). Given a sequence of events (①.  $\mathcal{E}$ ) and the corresponding questions (①.  $\mathcal{Q}$ ), EnigmaToM first leverages a chosen LLM (②) to identify key entities (e.g., characters and important objects) and their attributes relevant to ToM reasoning (Top left of ③). Enigma then produces state information for these entities after each event (Top right of ③, §3.1). With the entity state knowledge, EnigmaToM first conducts *Knowledge Injection* (referred to as "KI" thereafter) to enrich the original events by adding relevant fine-grained entity state details (§3.2). Among the entity state knowledge, spatial information of characters is used to conduct perspective-taking through an *Iterative Masking* mechanism (referred to as "IM" thereafter. Bottom of ③, §3.3). The modified events are provided to the LLM for final answers via zero-shot prompting (④). By offloading much of the ToM reasoning process to the symbolic IM component via perspective-taking, EnigmaToM reduces LLMs’ reasoning burden.

#### 3.1 The Enigma Neural Knowledge Base

NKBs such as COMET are trained on a large corpus of structured knowledge in a sequence-

to-sequence manner (Bosselut et al., 2019). Following this approach, we fine-tuned a Llama3.1-8B (Dubey et al., 2024) model to function as our NKB (Zheng et al., 2024).<sup>2</sup> For training, we used OpenPI2.0 (Zhang et al., 2024), which consists of 25,600 human-annotated entity state changes derived from WikiHow articles. OpenPI2.0 was selected over ATOMIC and Event2Mind as it contains more complex events and entity states. Alternatively, as LLMs become increasingly adept at commonsense reasoning, they can serve as an NKB of entity states via prompting (Hwang et al., 2021). We denote the trained (T) and prompt-based (P) NKB as Enigma<sup>T</sup> and Enigma<sup>P</sup>, respectively.

To query the NKB, we adopt an *entity-attribute-guided* approach which contains two steps. In Step 1, given a sequence of events,  $\mathcal{E}$ , a set of ToM questions,  $\mathcal{Q}$ , and a chosen LLM parameterized by  $\theta$ , we obtain a set of entities of interest,  $E = \{e\}_{i=1}^n$ , and their corresponding attributes,  $A = \{a\}_{j=1}^m$ , by zero-shot prompting:

$$E, A = \text{LLM}_\theta(\rho(\mathcal{E}, \mathcal{Q})) \quad (2)$$

where  $\rho$  denotes the prompt template (see Appendix D for details of the prompt). Then in Step 2, given an event,  $\epsilon \in \mathcal{E}$ , a set of entities of interest  $E$  and their corresponding attributes  $A$ , we query Enigma to retrieve the state of the entities after event  $\epsilon$ :

$$s_\epsilon = \bigoplus_{i=1}^n \bigoplus_{j=1}^m \text{Enigma}(e_i, a_j, \epsilon) \quad \forall \epsilon \in \mathcal{E}, e_i \in E, a_j \in A \quad (3)$$

where  $\oplus$  denotes concatenation.

### 3.2 Knowledge Injection (KI) with Enigma

In prior studies, perspective-taking was regarded as filtering out events unobserved by a given character, yielding a subset  $\mathcal{E}'_c \subseteq \mathcal{E}$ . We argue that beyond event filtering, perspective-taking should enhance LLMs' comprehension of events. Fine-grained entity state knowledge is crucial for event reasoning (Zhang et al., 2023) but often omitted due to reporting bias (Shwartz and Choi, 2020). To address this, we propose a knowledge injection mechanism, KI, that utilizes Enigma to enrich observable events with fine-grained entity state information. In the first step of KI, a chosen LLM is used to infer key entities,  $E$ , and attributes,  $A$ , based on a given sequence of events,  $\mathcal{E}$ , and a set of ToM questions,  $\mathcal{Q}$ ,

(Equation 2). We then query Enigma with the recognized entities and their attributes to obtain their state information at each event (Equation 3). We exclude spatial information of characters,  $\mathcal{S}_c^p$ , as this will be handled in the subsequent masking process (§3.3). Given a sequence of events,  $\mathcal{E} = \{\epsilon\}_{i=1}^n$ , we augment it by injecting entity state knowledge, resulting in the sequence  $\mathcal{E}^{\text{aug}}$ :

$$\mathcal{E}^{\text{aug}} = \bigoplus_{i=1}^n \epsilon_i \oplus \hat{s}_{\epsilon_i}, \text{ where } \hat{s}_{\epsilon_i} = s_{\epsilon_i} \setminus s_c^p \quad (4)$$

where  $\oplus$  denotes concatenation. As fine-grained entity state knowledge is often omitted in events due to reporting bias (Shwartz and Choi, 2020), this mechanism compensates for the lost information. More importantly, by providing state information of key entities, KI reinforces LLMs' understanding of the observed events.

### 3.3 Perspective-Taking (IM) with Enigma

Studies in psychology have shown that people's beliefs about others' mental states rely only on information available to themselves<sup>3</sup> (Arslan et al., 2017). Building on this insight, we assume that characters interpret others' beliefs through the lens of their own mental states, which allows us to employ *Iterative Masking* (IM) to facilitate efficient and accurate ToM reasoning across various order.

Perspective-taking with Enigma is accomplished by constructing spatial scene graphs and performing *Iterative Masking* (IM) using constructed graphs. Specifically, we obtain spatial information,  $\mathcal{S}_c^p$ , by querying Enigma about the location (*attr*) of a specific character (*ent*),  $c$ , using Equation (3). Spatial scene graphs are constructed based on spatial information to represent the detailed locations where each event takes place as perceived by a given character. The nodes of the scene graph represent events and locations, while the edges denote the "isin" relationship, specifying the location where each event takes place.

During IM, we first construct a character-oblivious spatial scene graph,  $G_{\mathcal{E}^{\text{aug}}}$ , which documents the *location* of each augmented event from an omniscient perspective. We then construct character-centric spatial scene graphs,  $G_c$ , that capture event locations from the perspective of each character. We introduce a null node,  $\emptyset$ , which indicates that the location of the current event is

<sup>2</sup>See Appendix B for details of the fine-tuning process.

<sup>3</sup>For instance, "Anne's belief about Sally's mental state" depends only on information available to Anne, i.e. events witnessed by Anne herself.



unknown to the character. During IM, the null node serves as a "mask" to exclude the event nodes, which are unobserved by the character, from  $G_{\mathcal{E}^{aug}}$  (see Figure 1 and Figure 2). For high-order ToM reasoning,  $G_{\mathcal{E}^{aug}}$  is masked sequentially by the order of characters in the belief chain<sup>4</sup>:

$$G_{c_{1:k}}^{masked} = G_{\mathcal{E}^{aug}} \bigotimes_{j=1}^k G_{c_j} \quad (5)$$

where  $\bigotimes$  represents the masking operation, and  $k$  corresponds to the ToM-order. The observable events of character  $c_{1:k}$  with injected entity state knowledge can be constructed as:

$$\mathcal{E}_{c_{1:k}}^{aug} = V_{G_{c_{1:k}}^{masked}}^{\epsilon} \quad (6)$$

where  $V_{G_{c_{1:k}}^{masked}}^{\epsilon}$  represents event nodes in  $G_{c_{1:k}}^{masked}$ . In the case of high-order ToM reasoning,  $\mathcal{E}_{c_{1:k}}^{aug}$  is obtained by iteratively applying the belief of characters. As such,  $\mathcal{E}_{c_{1:k}}^{aug}$  effectively encapsulates the beliefs of all characters in the belief chain. This allows us to transform the high-order ToM question to that of first-order. For instance, reasoning about "Sally's belief about Anne's belief" without EnigmaToM requires first inferring Sally's perceived world state, which then serves as the basis for modeling Anne's belief. With EnigmaToM, such nested dependencies and recursive reasoning are handled by the IM mechanism. Consequently, under  $\mathcal{E}_{Sally, Anne}^{aug}$ , deriving Sally's belief is sufficient to answer the original second-order Theory of Mind (ToM) question. Illustrative examples and further details on ToM order reduction are provided in Appendix C. We present illustrative examples and details of ToM order reduction in Appendix C.

### 3.4 Efficiency of EnigmaToM

The IM mechanism of EnigmaToM addresses the intractability of high-order ToM reasoning faced by SymbolicToM (Sclar et al., 2023). Due to the asymmetry of ToM modeling<sup>5</sup>, enumerating all possible mental states for characters across all ToM orders is a permutation problem. Suppose a ToM reasoning question involves  $m$  characters and the ToM order goes up to  $k^{th}$ -order, the worst-case complexity of constructing belief graphs in SymbolicToM is  $\mathcal{O}\left(\sum_{i=1}^k \frac{m!}{(m-i)!}\right)$ . In

<sup>4</sup>For instance, the masked spatial scene graph for "Sally's belief of Anne's mental state" is  $G_{\mathcal{E}^{aug}} \otimes G_{Sally} \otimes G_{Anne}$ .

<sup>5</sup>For example, in second-order ToM, Anne's belief of Sally's mental state is not equivalent to Sally's belief of Anne's mental state.

Dataset	O	Unit	#Units	#Qs
ToMi <sup>6</sup> (Le et al., 2019)	2	E	9.85	114
HiToM (Wu et al., 2023)	4	E	26.49	614
FANToM (Kim et al., 2023)	2	U	23.14	577

Table 1: Summary of datasets. **O**: highest ToM order tested. **Unit**: type of event sequence. "E" for event and "U" for utterance. **#Units**: avg. units per sequence. **#Qs**: avg. number of questions per sampled subset. Examples from each dataset can be found in Appendix A.

contrast, EnigmaToM constructs one spatial scene graph,  $G_{\mathcal{E}^{aug}}$ , which encapsulates omniscient spatial information, and  $m$  character-centric spatial scene graphs. Hence, the worst-case complexity for constructing spatial scene graphs in EnigmaToM is  $\tilde{T}(m, k) = \mathcal{O}(m)$ , which is linear with respect to the number of characters and independent of the ToM order  $k$ . We illustrate the difference in complexity in Appendix E.

## 4 Experiments

EnigmaToM is evaluated on three widely used ToM benchmarks (Table 1) and compared against the following generic and ToM-specific methods:

**CoT** (Wei et al., 2022) boosts LLMs' reasoning capabilities by prompting LLMs to explicitly list out their reasoning process.

**SimToM** (Wilf et al., 2024) conducts perspective-taking by directly querying the LLMs about the mental states of characters.

**TimeToM<sup>†</sup>** (Hou et al., 2024) leverage the temporal information of events to conduct perspective-taking. The final answer is obtained using a multi-perspective belief-solving prompt.

**DWM** (Huang et al., 2024) conducts perspective-taking by partitioning the events into chunks and querying the LLMs about characters' mental states after each chunk.

**PerceptToM<sup>†</sup>** (Jung et al., 2024) conducts perspective-taking by querying the LLMs about the characters' awareness of the events.

To ensure a fair comparison with established methods, we conduct controlled experiments by controlling the format and answer space of all ToM questions. In addition, we follow a realistic setting of ToM reasoning by using only the sequence of events and ToM questions from each dataset.

<sup>6</sup>We use the disambiguated ToMi (Sclar et al., 2023) from <https://github.com/msclar/symbolictom>.

<sup>†</sup>Official implementation is not available at the time of experiments (Sept-Dec, 2024). We implemented this method using prompts from the corresponding paper.

		Qwen2.5-7B	Llama3.1-8B	Gemma2-9B	Gemma2-27B	Llama3.3-70B <sup>4bit</sup>	Qwen2.5-72B <sup>4bit</sup>	GPT-4o
ToMi	Vanilla	0.722±0.045	0.647±0.011	0.741±0.037	0.715±0.048	0.767±0.015	0.717±0.034	0.767±0.041
	CoT	0.724±0.026	0.739±0.025	0.676±0.035	0.537±0.056	0.741±0.032	0.767±0.033	0.769±0.029
	SimToM	0.642±0.022	0.600±0.020	0.710±0.034	0.684±0.015	0.712±0.018	0.749±0.020	0.749±0.018
	TimeToM	0.567±0.024	0.630±0.019	0.681±0.028	0.587±0.036	0.739±0.021	<b>0.865±0.018</b>	0.723±0.016
	DWM	0.686±0.023	0.644±0.033	0.718±0.028	0.707±0.045	0.735±0.016	0.762±0.051	0.739±0.049
	PerceptToM	0.720±0.038	0.695±0.025	0.676±0.029	0.749±0.017	0.738±0.032	0.809±0.033	0.790±0.023
	Enigma <sup>p</sup>	0.706±0.044	0.738±0.056	<b>0.865±0.031</b>	<b>0.833±0.018</b>	<b>0.828±0.012</b>	0.839±0.014	<b>0.847±0.030</b>
	Enigma <sup>t</sup>	<b>0.825±0.030</b>	<b>0.796±0.023</b>	0.814±0.020	0.804±0.050	0.787±0.024	0.837±0.024	0.795±0.036
HiToM	Vanilla	0.378±0.013	0.333±0.015	0.471±0.009	0.527±0.018	0.534±0.008	0.456±0.012	0.521±0.006
	CoT	0.441±0.007	0.304±0.021	0.474±0.008	0.535±0.018	0.537±0.011	0.481±0.011	0.527±0.005
	SimToM	0.402±0.009	0.368±0.024	0.473±0.012	0.549±0.018	0.569±0.005	0.536±0.018	0.571±0.003
	TimeToM	0.316±0.010	0.462±0.013	0.302±0.012	0.302±0.013	0.623±0.006	0.415±0.013	0.633±0.008
	DWM	0.444±0.020	0.367±0.019	0.485±0.012	0.488±0.018	0.564±0.010	0.560±0.009	0.580±0.018
	PerceptToM	0.393±0.019	0.342±0.011	0.440±0.009	0.562±0.007	0.588±0.010	0.548±0.016	0.580±0.018
	Enigma <sup>p</sup>	<b>0.508±0.012</b>	<b>0.477±0.005</b>	<b>0.555±0.010</b>	<b>0.576±0.004</b>	<b>0.696±0.007</b>	<b>0.605±0.007</b>	<b>0.733±0.017</b>
	Enigma <sup>t</sup>	0.457±0.005	0.431±0.010	0.446±0.008	0.478±0.004	0.518±0.011	0.473±0.010	0.626±0.020
FANToM	Vanilla	0.400±0.015	0.429±0.022	0.485±0.016	0.553±0.011	0.486±0.022	0.532±0.025	0.476±0.020
	CoT	0.398±0.014	0.438±0.014	0.470±0.019	0.556±0.007	0.494±0.028	0.521±0.024	0.453±0.014
	SimToM	0.413±0.012	0.440±0.015	0.427±0.009	0.574±0.010	<b>0.620±0.025</b>	0.516±0.014	0.502±0.016
	TimeToM	0.252±0.020	0.260±0.012	0.299±0.011	0.300±0.021	0.580±0.017	0.409±0.026	0.404±0.016
	DWM	0.429±0.013	0.470±0.027	0.433±0.023	0.562±0.017	0.473±0.021	0.543±0.014	0.465±0.028
	PerceptToM	0.408±0.023	0.407±0.026	0.504±0.006	<b>0.611±0.009</b>	0.527±0.011	0.573±0.016	0.521±0.006
	Enigma <sup>p</sup>	0.445±0.026	0.442±0.018	0.439±0.023	0.462±0.014	0.515±0.020	0.450±0.013	0.531±0.015
	Enigma <sup>t</sup>	<b>0.487±0.018</b>	<b>0.545±0.036</b>	<b>0.530±0.012</b>	0.582±0.028	0.610±0.021	<b>0.574±0.031</b>	<b>0.553±0.011</b>

Table 2: Main results of EnigmaToM in comparison with existing methods on ToMi, HiToM, and FANToM datasets. Accuracy means and variances are calculated based on 5 runs, which used 5 different subsets of the corresponding dataset. The best and second best results are highlighted in **bold** and underline respectively.

Auxiliary information such as character names is obtained using an off-the-shelf NER model<sup>8</sup>.

**Question Formatting** We formulate ToMi as a free-form generation task where the model is instructed to choose between two possible answers. We formulate HiToM as a multiple-choice task as in the original paper (Wu et al., 2023). FANToM contains both free-form generation and multiple-choice questions. We follow the question formatting instructions in the original paper (Kim et al., 2023). For efficient and accurate parsing of LLM responses, we follow the convention of (Huang et al., 2024), instructing LLMs to wrap answers within the special <answer> and </answer> tokens. As introduced in §3.3, the recursive modeling of mental states in high-order ToM questions has been addressed by the IM mechanism, which allows us to transform high-order ToM questions into first-order questions. Similarly, TimeToM leverages temporal information to conduct symbolic modeling of high-order ToM (Hou et al., 2024). We apply such transformation when evaluating with TimeToM and EnigmaToM (Appendix C).

<sup>8</sup><https://huggingface.co/dslim/bert-large-NER>

**Towards Robust Evaluation** To ensure robust evaluation, we construct 5 subsets for each dataset by sampling data points using commonly used random seeds<sup>†</sup>. Each subset of ToMi and HiToM contains 100 event sequences, whereas each subset of FANToM contains 50 multi-round dialogues. The number of QA pairs in each subset is shown in Table 1. We report both the mean accuracy and its variance based on the 5 runs.


We evaluate each method using various instruction-tuned LLMs, including Llama3.1-8B, Llama3.3-70B<sup>‡</sup>, Qwen2.5-7B, Qwen2.5-72B<sup>‡</sup>, Gemma2-9B, Gemma2-27B, and GPT-4o (Dubey et al., 2024; Yang et al., 2024; Gemma, 2024; OpenAI, 2024). To ensure reproducibility, all experiments are done using zero-shot prompting with greedy decoding and a temperature of 0. LLM inference is carried out using LLM on 2 NVIDIA A100<sup>80GB</sup> GPUs (Kwon et al., 2023).

Table 2 shows the main results of EnigmaToM in comparison with existing methods on ToMi,

<sup>†</sup>We use 12, 42, 96, 2012, and 2024 as random seeds.

<sup>‡</sup>Loaded in 4bit using BitsandBytes (Dettmers et al.) with weights from <https://huggingface.co/unsloth>.

HiToM, and FANToM datasets. In general, we see that EnigmaToM brings improvements in accuracy across all datasets and most LLMs. Specifically, Enigma<sup>P</sup> outperforms other methods on ToMi and HiToM, while Enigma<sup>T</sup> achieves superior performance on FANToM. EnigmaToM is particularly effective with smaller LLMs. For instance, Enigma<sup>T</sup> boosts Qwen2.5-7B to exceed the zero-shot performance of Qwen2.5-72B<sup>4bit</sup>. Further, results from the HiToM dataset demonstrate that EnigmaToM is particularly effective in high-order ToM reasoning. We analyze the effectiveness of EnigmaToM in tackling high-order ToM reasoning in §5.1. Moreover, results from Table 2 show that Enigma<sup>P</sup> performs better on event-based datasets (ToMi and HiToM) while Enigma<sup>T</sup> is more effective on a dialogue-based dataset (FANToM). We investigate such a discrepancy in §5.2 and §5.3.

## 5 Analysis

### 5.1 High-Order ToM Reasoning

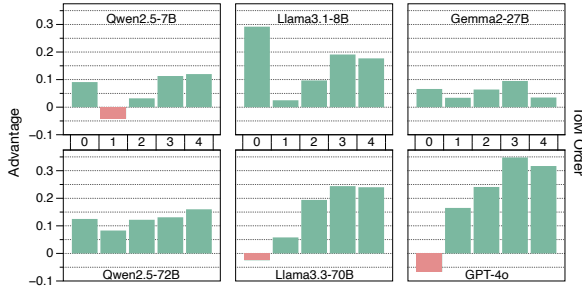


Figure 3: Relative advantage of EnigmaToM on HiToM dataset with respect to ToM order.

To assess the effectiveness of EnigmaToM in high order ToM reasoning, we analyze its performance on the HiToM dataset, which consists of ToM questions requiring reasoning up to the fourth order. We compute the relative advantage of EnigmaToM with Enigma<sup>P</sup> over the zero-shot vanilla prompting baseline. From Figure 3, we observe that EnigmaToM improves mean accuracy across all orders of ToM reasoning, with notable effectiveness in higher-order ToM reasoning. Specifically, results from the Qwen2.5 and Llama3 families demonstrate that EnigmaToM has an increasing advantage as the order of ToM reasoning increases. For the third- and fourth-order ToM reasoning, EnigmaToM achieves an average improvement of  $0.160 \pm 0.003$  and  $0.148 \pm 0.004$  respectively, across all models compared to the baseline. We observe similar trends on ToMi and FANToM albeit they only contain ToM questions up to the second

		Llama3.3-70B <sup>4bit</sup>	Qwen2.5-72B <sup>4bit</sup>	GPT-4o
ToMi	Enigma <sup>P</sup>	0.828 $\pm$ 0.012	0.839 $\pm$ 0.014	0.847 $\pm$ 0.030
	Enigma <sup>T</sup>	0.787 $\pm$ 0.024	0.837 $\pm$ 0.024	0.795 $\pm$ 0.036
	w/o KI	0.834 $\pm$ 0.067	0.845 $\pm$ 0.026	0.811 $\pm$ 0.028
	w/o IM	0.693 $\pm$ 0.014	0.655 $\pm$ 0.039	0.674 $\pm$ 0.002
	w/o KI, IM	0.767 $\pm$ 0.015	0.717 $\pm$ 0.034	0.767 $\pm$ 0.041
HiToM	Enigma <sup>P</sup>	0.696 $\pm$ 0.007	0.605 $\pm$ 0.007	0.733 $\pm$ 0.017
	Enigma <sup>T</sup>	0.518 $\pm$ 0.011	0.473 $\pm$ 0.010	0.626 $\pm$ 0.020
	w/o KI	0.726 $\pm$ 0.004	0.632 $\pm$ 0.003	0.751 $\pm$ 0.004
	w/o IM	0.460 $\pm$ 0.013	0.423 $\pm$ 0.008	0.442 $\pm$ 0.006
	w/o KI, IM	0.534 $\pm$ 0.008	0.456 $\pm$ 0.012	0.521 $\pm$ 0.006
FANToM	Enigma <sup>P</sup>	0.515 $\pm$ 0.020	0.450 $\pm$ 0.013	0.531 $\pm$ 0.015
	Enigma <sup>T</sup>	0.610 $\pm$ 0.021	0.574 $\pm$ 0.031	0.553 $\pm$ 0.011
	w/o KI	0.607 $\pm$ 0.018	0.542 $\pm$ 0.036	0.539 $\pm$ 0.012
	w/o IM	0.500 $\pm$ 0.021	0.477 $\pm$ 0.017	0.470 $\pm$ 0.013
	w/o KI, IM	0.486 $\pm$ 0.002	0.532 $\pm$ 0.025	0.476 $\pm$ 0.020

Table 3: Ablation study of EnigmaToM on ToMi, HiToM, and FANToM datasets. "w/o KI" indicates without *entity state knowledge injection*. "w/o IM" denotes without *perspective-taking via iterative masking*. ● Improved and ● decreased results are highlighted.

order. See Appendix F for complete results and analysis on all three datasets.

### 5.2 Ablation Study

To understand the effectiveness of each component of EnigmaToM, we conduct an ablation study by (1) keeping the injected knowledge but removing the masking-based perspective-taking mechanism (directly using  $\mathcal{E}^{\text{aug}}$  as context); and (2) conducting perspective-taking without knowledge injection (applying Equation 5 with  $G_{\mathcal{E}}$  instead of  $G_{\mathcal{E}^{\text{aug}}}$ ).

**Enigma for Perspective Taking** As shown in Table 3, removing the IM mechanism results in an average accuracy drop of  $-0.165$  on ToMi,  $-0.172$  on HiToM, and  $-0.103$  on FANToM. This suggests that the iterative masking mechanism is effective in perspective-taking and crucial for EnigmaToM to achieve boosted performance in ToM reasoning (See Table A2 for complete results).

**Enigma for Knowledge Injection** Compared to IM for perspective-taking, entity state knowledge injection is less critical. On ToMi and HiToM, its removal slightly reduces Gemma2-27B’s performance on ToMi but improves performance for all other LLMs on both benchmarks, further highlighting IM’s effectiveness in perspective-taking. However, for FANToM, entity state knowledge is indispensable, as excluding it results in performance drops across all LLMs. For ToMi and HiToM, we hypothesize that larger LLMs are better at handling reporting bias. This aligns with the results shown in Table 2, where Enigma<sup>P</sup> surpasses Enigma<sup>T</sup> as LLM

Dataset	Model	Precision	Recall	F1-Score
TMi	Llama3.3-70B <sup>4bit</sup>	0.859	0.968	0.910
FTM	Llama3.3-70B <sup>4bit</sup>	0.880	0.970	0.923

Table 4: Performance analysis of key entity recognition in ToMi (TMi) and FANToM (FTM) datasets using Llama3.3-70B<sup>4bit</sup>. See Appendix J for detailed description of the evaluation process.

	Model	Relevance	Accuracy	Avg. #Token
TMi	Enigma <sup>T</sup> <sub>8B</sub>	0.847	0.807	7.665
	Enigma <sup>T</sup> <sub>70B</sub>	0.870	0.860	9.740
FTM	Enigma <sup>T</sup> <sub>8B</sub>	0.880	0.773	8.973
	Enigma <sup>T</sup> <sub>70B</sub>	0.880	0.700	30.517

Table 5: Performance analysis of Enigma<sup>T</sup> on ToMi (TMi) and FANToM (FTM) datasets. See Appendix J for detailed description of the evaluation process.

size increases, meaning that the fine-grained information about the state of the entity and its causal relationships with events are encapsulated more effectively in the larger LLMs. In such cases, potential inaccuracies in injected entity-state knowledge outweigh its benefits in addressing reporting bias, leading to decreased performance. In the case of FANToM, the dialogue-based nature of the dataset makes useful information sparser than in event-based datasets. Here, knowledge injection serves a different role: rather than primarily addressing reporting bias, it compresses important information from utterances into entity-state representation, effectively reducing LLMs’ workload in identifying crucial information. See Appendix H for examples.

### 5.3 Effectiveness of LLMs and Enigma

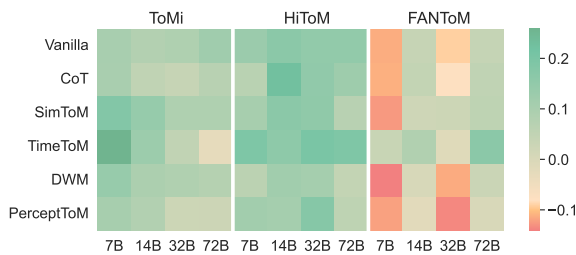


Figure 4: Relative advantage of EnigmaToM on ToMi, HiToM, and FANToM datasets. We use Enigma<sup>P</sup> as the pivot method for ToMi and HiToM and Enigma<sup>T</sup> for FANToM. Exact mean accuracies are shown in Table A1. Model sizes shown in x-axis are Qwen2.5 models.

The effectiveness of EnigmaToM is contingent upon the capabilities of both the Enigma neural knowledge base and the LLM deployed in the framework. In this section, we conduct analysis

of EnigmaToM with an aim to explore the following two questions: (1) Does EnigmaToM benefit LLMs of larger size? and (2) How effective is our Enigmeneural knowledge base and does scaling Enigmalead to increased performance?

We raise the first question by hypothesizing that perspective-taking, albeit the challenges posed by its multi-hop nature, could become solvable by more capable LLMs. Empirically speaking, the capability of LLMs positively correlates to their number of parameters. To eliminate potential confounding factors, we analyze the effectiveness of LLMs from the Qwen2.5 family with sizes ranging from 7B to 72B (Yang et al., 2024).

In the second question, we aim to examine the effectiveness of Enigma<sup>T</sup>. Trained using data from OpenPI2.0, we wish to investigate how well can the knowledge encapsulated in Enigma<sup>T</sup> be transferred to aid ToM reasoning. Aside from the performance of Enigma<sup>T</sup>, we also explore the effectiveness of scaling of Enigma<sup>T</sup>. In addition to the Enigma<sup>T</sup> used in previous experiments, which is trained using a Llama3.1-8B model, we trained another Enigma<sup>T</sup> based on Llama3.3-70B, which we denote as Enigma<sup>T</sup><sub>70B</sub>. Experiments with Enigma<sup>T</sup><sub>70B</sub> are carried out following the same procedure described in §4.

**Scaling of Base LLMs** We compute the relative advantage of EnigmaToM by calculating the difference in mean accuracy between EnigmaToM and the most performant baseline methods (see Table 2). We use Enigma<sup>P</sup> as the pivot method for ToMi and HiToM and Enigma<sup>T</sup> for FANToM. Figure 4 shows two trends: (1) a slight diminishment in advantage on ToMi and (2) a gradual increase in advantage on FANToM. We attribute this to the differing difficulty levels of these two datasets. ToMi, which consists of short sequences of concise events, becomes easier to solve with large-scale LLMs. Conversely, FANToM, featuring long sequences of lengthy dialogues, remains challenging even for larger LLMs. HiToM, positioned between these two extremes with long sequences of concise events, shows that EnigmaToM has a consistent advantage regardless of the LLM sizes. This discrepancy in performance and model scaling effect between ToMi and FANToM aligns with the analysis in §4 and §5.3. These findings suggest that while prompting large-scale LLMs can potentially tackle ToM reasoning involving short event sequences (as in ToMi), ToM reasoning about lengthy event or dialogue sequences (as in HiToM and FANToM) can benefit from the



		Qwen2.5-7B	Llama3.1-8B	Gemma2-9B	Gemma2-27B	Llama3.3-70B <sup>4bit</sup>	Qwen2.5-72B <sup>4bit</sup>	GPT-4o
ToMi	Enigma <sup>P</sup>	0.706 $\pm$ 0.044	0.738 $\pm$ 0.056	0.865 $\pm$ 0.031	0.833 $\pm$ 0.018	0.828 $\pm$ 0.012	0.839 $\pm$ 0.014	0.847 $\pm$ 0.030
	Enigma <sup>T</sup> <sub>8B</sub>	0.825 $\pm$ 0.030	0.796 $\pm$ 0.023	0.814 $\pm$ 0.020	0.804 $\pm$ 0.050	0.787 $\pm$ 0.024	0.837 $\pm$ 0.024	0.795 $\pm$ 0.036
	Enigma <sup>T</sup> <sub>70B</sub>	0.837 $\pm$ 0.017	0.835 $\pm$ 0.026	0.847 $\pm$ 0.008	0.854 $\pm$ 0.023	0.844 $\pm$ 0.018	0.860 $\pm$ 0.015	0.872 $\pm$ 0.026
HiToM	Enigma <sup>P</sup>	0.508 $\pm$ 0.012	0.477 $\pm$ 0.005	0.555 $\pm$ 0.010	0.576 $\pm$ 0.004	0.696 $\pm$ 0.007	0.605 $\pm$ 0.007	0.733 $\pm$ 0.017
	Enigma <sup>T</sup> <sub>8B</sub>	0.457 $\pm$ 0.005	0.431 $\pm$ 0.010	0.446 $\pm$ 0.008	0.478 $\pm$ 0.004	0.518 $\pm$ 0.011	0.473 $\pm$ 0.010	0.626 $\pm$ 0.020
	Enigma <sup>T</sup> <sub>70B</sub>	0.456 $\pm$ 0.007	0.444 $\pm$ 0.012	0.481 $\pm$ 0.006	0.489 $\pm$ 0.012	0.517 $\pm$ 0.010	0.467 $\pm$ 0.009	0.733 $\pm$ 0.010
FANToM	Enigma <sup>P</sup>	0.445 $\pm$ 0.026	0.442 $\pm$ 0.018	0.439 $\pm$ 0.023	0.462 $\pm$ 0.014	0.515 $\pm$ 0.020	0.450 $\pm$ 0.013	0.531 $\pm$ 0.015
	Enigma <sup>T</sup> <sub>8B</sub>	0.487 $\pm$ 0.018	0.545 $\pm$ 0.036	0.530 $\pm$ 0.012	0.582 $\pm$ 0.028	0.610 $\pm$ 0.021	0.574 $\pm$ 0.031	0.553 $\pm$ 0.011
	Enigma <sup>T</sup> <sub>70B</sub>	0.479 $\pm$ 0.032	0.517 $\pm$ 0.041	0.491 $\pm$ 0.039	0.529 $\pm$ 0.028	0.537 $\pm$ 0.011	0.494 $\pm$ 0.023	0.539 $\pm$ 0.012

Table 6: Results of scaling Enigma<sup>T</sup> from Llama3.1-8B to Llama3.3-70B. ● Improved, ● Unchanged, and ● decreased results are highlighted in the corresponding color.

fine-grained entity state knowledge as well as the symbolic masking mechanism of EnigmaToM.

**Effectiveness of LLMs in Recognizing Key Entity-Attributes** Recognizing entities and their attributes (Equation 2) that are indispensable for answering the ToM questions posed is a critical pre-requisite for the effectiveness of EnigmaToM. On the one hand, failing to recognize a key entity will disable EnigmaToM to properly augment the events with critical information. On the other hand, erroneously identify an extraneous entity will lead to inclusion of redundant information, which will prolong the context and increase the reasoning burden of the LLM. To evaluate the quality of key entities and attributes extracted by LLMs, we manually labeled 300 entities and attributes identified using Llama3.3-70B<sup>4bit</sup>. Evaluation results from Table 4 suggests that LLMs are more than competent in identifying key entities and attributes. With a F1-score of 0.910 on the ToMi dataset and 0.923 on the FANToM dataset, it is safe to conclude that the vast majority of entities identified by LLMs are indeed vital to answering the ToM questions posed.

**Effectiveness of Enigma in Generating Entity State Information** To understand the effectiveness of scaling Enigma, we trained two Enigma<sup>T</sup> models using the same OpenPI2.0 dataset. With Llama3.1-8B as the base model, we trained Enigma<sup>T</sup><sub>8B</sub>. Further, with Llama3.3-70B, we trained Enigma<sup>T</sup><sub>70B</sub><sup>11</sup>. Table 6 shows that there is an obvious discrepancy between the scaling effect of Enigma<sup>T</sup>: Enigma<sup>T</sup><sub>70B</sub> consistently outperforms Enigma<sup>T</sup><sub>8B</sub> in ToMi and HiToM while underperforming Enigma<sup>T</sup><sub>8B</sub> in FANToM.

- **Relevance:** whether the entity and attribute contribute to answering the ToM question. This evaluates the same aspects of EnigmaToM as the precision scores shown in Table 4

- **Accuracy:** whether the entity state can be inferred from the given context.

From Table 5, we see that the relevance scores of both ToMi and FANToM exceed 80%, indicating that Llama3.3-70B is capable of identifying entities and attributes useful for ToM reasoning. Entity state length and accuracy are closely correlated. Enigma<sup>T</sup><sub>70B</sub> produces a more articulated response compared to its counterpart. Specifically, scaling Enigma<sup>T</sup> brings 5.3% improvement in accuracy on ToMi while increasing response length by only 2.075 tokens. In contrast, FANToM experiences a significant 21.544 token increase in response length, which reduces Enigma<sup>T</sup>’s efficiency as an information compressor and leads to greater hallucination, resulting in a 7.3% drop in accuracy. We provide demonstration examples in Appendix I.

## 6 Conclusion

In this work, we introduced EnigmaToM, a neuro-symbolic framework designed to enhance the ToM reasoning capabilities of LLMs. By leveraging a Neural Knowledge Base of Entity States through an iterative masking mechanism and knowledge injection, EnigmaToM accomplishes the bulk of ToM reasoning via perspective-taking through symbolic reasoning, which alleviates LLMs’ reasoning burden. Experimental results across multiple benchmarks demonstrate that EnigmaToM outperforms existing methods, particularly excelling in high-order ToM reasoning scenarios. Our analysis highlights the effectiveness of the iterative masking mechanism in maintaining strong performance across varying depths of ToM reasoning, as well as the critical role of fine-grained entity state knowledge in compressing key information in complex event sequences (as in FANToM). Furthermore, the framework’s efficiency and scalability make it a promising solution for addressing the computational challenges associated with high-order ToM reasoning tasks.

<sup>11</sup>Training details are provided in Appendix B.

## Limitations

### ToM Reasoning Beyond Character Perception

EnigmaToM tackles ToM reasoning of characters' beliefs based on their perceptions. While we believe that reasoning about characters' perceptions serve as a cornerstone for all types of ToM reasoning, future work may explore methods to facilitate real-world ToM reasoning about characters' emotions, intentions, desires, and their inherent subjectivity (Zhou et al., 2025).

**Neural Knowledge Base** EnigmaToM relies on access to a Neural Knowledge Base (NKB) to retrieve entity-state information for answering ToM questions. While Table 5 shows that Enigma is capable of producing accurate entity-state information, it can be further improved (e.g. full-parameter fine-tuning instead of LoRA). Further, expanding the NKB to incorporate richer entity-state details, including emotional, temporal, and causal relationships, would be beneficial for ToM reasoning about high-level information.

**Error Propagation** While experiments demonstrate the effectiveness of the IM mechanism, it is prone to error propagation. In the case of high-order ToM reasoning, applying a wrong mask in the iterative masking process will lead to the event being erroneously excluded and vice versa. Additionally, in cases requiring complex reasoning about non-linear or intertwined event dependencies, the symbolic Iterative Masking (IM) mechanism may need to be enhanced.

## Ethics Statement

This study aims to enhance LLMs' ToM reasoning by improving the accuracy and efficiency of perceptual perspective-taking, ultimately optimizing their effectiveness in communication. ToM reasoning is essential for enhancing LLMs' ability to interact with humans (e.g., in chatbots) or other LLMs (e.g., in multi-agent systems). The evaluation datasets used in this study have been peer-reviewed and widely adopted in previous research. However, these datasets may introduce issues such as cultural bias and often lack demographic information. Future research could incorporate auxiliary data, such as demographic and personality traits, to improve representativeness across diverse ethnic and cultural backgrounds.

## Acknowledgments

This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through an iCASE award with Huawei London Research Centre and a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2).

## References

- Ian Apperly. 2010. *Mindreaders: the cognitive basis of "theory of mind"*. Psychology Press.
- Burcu Arslan, Niels A Taatgen, and Rineke Verbrugge. 2017. Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: a computational modeling study. *Frontiers in psychology*, 8:275.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. **COMET: Commonsense transformers for automatic knowledge graph construction**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- 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)*, pages 15959–15983, Bangkok, Thailand. 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.
- Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. 2017. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, 31:250–287.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.