Beyond Knowing and Not Knowing: Tracking Structured Ignorance for Narrative Role-Playing

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used for narrative role-playing, yet they frequently break character by knowing too much, revealing information prematurely, or inferring hidden facts from an omniscient perspective. We argue that faithful role-play requires explicit modeling of structured ignorance rather than knowledge alone, since a believable character is defined as much by what they do not know, suspect, or cannot say as by what they know. In this work, we introduce the *Ignorance Graph*, a plot-indexed representation that tracks each character's stance toward narrative propositions throughout a story. The graph factorizes internal epistemic state into four main labels, UNAWARE, SUSPECTING, MISBELIEVING, and KNOWN, and further attaches one optional expression-level modifier, either WITHHELD or INEXPRESSIBLE, when a character's verbalization is constrained. Building on temporally ordered novels, we construct a benchmark consisting of two tasks: ignorance state classification and state transition prediction. Experiments with general-purpose LLMs reveal that structured ignorance remains a challenging problem. The strongest model reaches only 64.7 macro-F_1 on Task 1 and 68.4 macro-F_1 on Task 2, and optional expression modifiers prove substantially harder than main-state prediction. These results suggest that current LLMs lack reliable mechanisms for reasoning about character-bound epistemic states, pointing to a concrete direction for improving role-play fidelity.

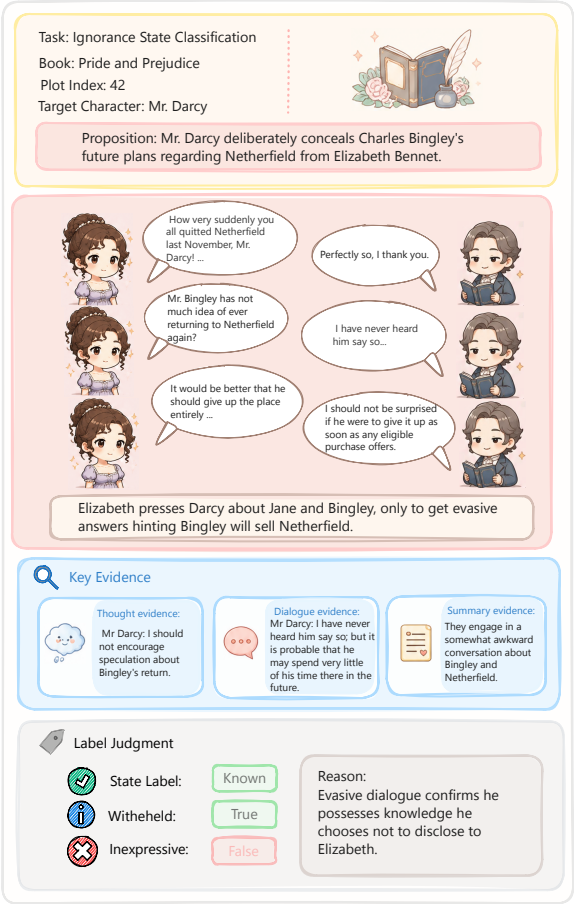


Figure 1: IgnoranceBench extracts character-specific propositions from narratives to track dynamic epistemic states. The dataset annotates unawareness, suspicion, misbelief, and knowledge alongside expression constraints like withheld information and inexpressibility.

1 Introduction

Large language models (LLMs) are increasingly deployed as role-playing agents that simulate characters, professions, and social relationships across extended dialogue. Role-play provides a useful abstraction for describing agent behavior without anthropomorphism (Shanahan et al., 2023), while benchmarks such as RoleLLM (Wang et al., 2024a) catalog how different prompting and tuning pipelines affect persona fidelity.

Recent evaluations probe role-play along new axes but still leave gaps that matter for narrative safety and immersion. DMT-RoleBench measures multi-turn dialogue traits, ESC-Eval targets emotion support roles, RVBench audits value alignment, and Role-Playing Eval introduces au-

tomated persona grading (Yuan et al., 2025; Zhao et al., 2024; Wang et al., 2025b; El Boudouri et al., 2025). However, these resources primarily focus on dialogue quality, emotional support, value alignment, or persona consistency, rather than on whether characters respect narrative knowledge boundaries, such as what they know, do not know, suspect, falsely believe, conceal, or struggle to articulate at a given point in the story.

Epistemically oriented benchmarks have begun to emerge. Reflection-Bench examines cognitive agency, but it primarily targets general reasoning rather than narrative consistency with a character's knowledge boundaries (Li et al., 2024). Narrative-evaluation work such as NarraBench further shows that existing datasets provide limited coverage of perspective-taking and revelation, highlighting the need for evaluations that track plot progression while enforcing character-limited viewpoints (Hamilton et al., 2026).

We address this gap with the *Ignorance Graph*, a plot-indexed representation that factorizes internal epistemic states and expression constraints for every character-proposition pair in narrative role-play scenarios. Using temporally ordered novels, we construct two benchmark tasks: ignorance state classification and state-transition prediction, each exposing distinct failure modes when a model collapses ignorance. Figure 1 illustrates the construction of our benchmark from the narrative data of Wang et al. (2025a). Building on this resource, our benchmark identifies a target character within a specific plot and extracts character-specific propositions, each assigned a primary epistemic state and an optional expression modifier supported by evidence from the narrative context.

Our contributions are:

- We introduce the *Ignorance Graph*, an ontology that decouples four internal epistemic states (UNAWARE, SUSPECTING, KNOWN, MISBELIEVING) from one optional expression constraint (WITHHELD or INEXPRESSIBLE) to evaluate character knowledge boundaries, false beliefs, and pragmatic concealment.
- We construct **IgnoranceBench** by extracting atomic propositions and inferring character states across sequential novels to formulate ignorance state classification and state transition prediction tasks.
- We evaluate ten LLMs and show that structured ignorance remains unsolved. Even leading models such as GPT-5.4 exhibit leakage errors, while open-weight systems struggle with fine-grained state distinctions.

2 Related Work

2.1 Role-Playing Evaluation for Large Language Models

The evaluation of LLM role-playing has moved from basic persona consistency checks to comprehensive character fidelity benchmarks. Early research by Shao et al. (2023) and Zhou et al. (2023) established the foundation of character simulation. Wang et al. (2024a) and Chen et al. (2024a) subsequently formalized this problem through systematic evaluation pipelines that test character-specific knowledge, linguistic style, and behavioral patterns. Building on these foundations, later benchmarks expanded the evaluation scope across complementary dimensions. Tu et al. (2024) and Wang et al. (2025a) introduced multi-dimensional assessments using literary data to evaluate established fictional personas. Concurrently, Samuel et al. (2024) and Wang et al. (2024b) examined personality fidelity through psychometric and dialogue-based protocols. Furthermore, Kong et al. (2024) demonstrated that perspective-taking serves as a beneficial inductive bias, showing that role-play prompting can enhance general zero-shot reasoning.

2.2 Epistemic State and Theory of Mind Reasoning

A parallel research trajectory investigates theory-of-mind reasoning to determine whether LLMs can accurately represent the beliefs of others. Initial benchmarks identified significant deficiencies in belief tracking among standard models (Kim et al., 2023; Wu et al., 2023). However, explicit perspective-taking and dedicated belief trackers have been shown to substantially improve performance (Sclar et al., 2023; Kosinski, 2024; Wilf et al., 2024). Recent evaluations highlight that this apparent competence remains vulnerable to adversarial perturbations (Shapira et al., 2024) and requires further validation in interactive opponent modeling (Kuru et al., 2025). Closely related to our definition of character-specific epistemic states, Li et al. (2024) evaluated belief updating and metacognitive reflection. Similarly,

Mireshghallah et al. (2024) investigated contextual integrity and secret-keeping. This line of work aligns with the WITHHELD state in our ontology, although it focuses on single-turn privacy disclosure rather than narrative role-play.

2.3 Narrative Understanding and Proposition-Level Modeling

Narrative understanding requires reasoning about temporality, information asymmetry, and character-specific mental states. Foundational studies formalized this domain through naive psychology modeling (Rashkin et al., 2018), world state tracking (Henaff et al., 2016), and reading comprehension frameworks (Kočiskỳ et al., 2018; Lei et al., 2020; Brahman et al., 2021). More recently, Hamilton et al. (2026) emphasized the need to evaluate discourse structure and point of view to capture suspense and information disclosure. From a theoretical standpoint, Shanahan et al. (2023) conceptualized LLM interactions as character simulations rather than direct expressions of model beliefs, providing a basis for fictional role-playing under limited perspectives. To represent narrative content at a fine granularity, researchers have developed proposition-level segmentation (Chen et al., 2023, 2024c) and atomic-claim verification frameworks (Min et al., 2023; Kamoi et al., 2023). These technical advances enable the decomposition of narratives into discrete facts and facilitate the verification of entailment relations, which directly supports the construction of our Ignorance Graph.

3 Approach

We introduce **IgnoranceBench**, a benchmark for evaluating whether LLMs can maintain *structured ignorance* when role-playing characters in long-form narratives, i.e., explicitly modeling *what a character does not know, mistakenly disbelieves, deliberately conceals, or cannot articulate*. The benchmark is built upon CoSER (Wang et al., 2025a), which provides processed narrative materials from novels, including chapter-level plot summaries, character dialogues, inner thoughts, physical actions, character experiences, and conversational settings. From this resource, we retain novels satisfying three criteria: (i) strict chronological narration, (ii) complete causal transformations, and (iii) rich and nuanced emotional depth of characters, yielding 17 novels and two task-specific evaluation sets.

Our construction pipeline consists of two phases: (i) proposition extraction and normalization (§ 3.1.1), and (ii) ignorance state inference and graph construction (§ 3.1.2). Table 1 summarizes the resulting benchmark scale and gold-label counts across 17 novels and two evaluation tasks. IgnoranceBench consists of 1,318 propositions and 18,986 character-proposition-time triples as the full dataset. After downsampling for unaware and known status classification, we obtain 3,132 current-state samples and 1,741 transition-prediction samples.

Figure 2 summarizes the end-to-end construction pipeline. The procedure is designed to be reproducible and auditable. It converts narrative text into proposition-level units, aligns them with character-centric evidence, and then propagates epistemic states over time to form the final graph representation.

3.1 Benchmark Construction

Let a novel be segmented into a temporally ordered sequence of chapters $\mathcal{C} = (c_1, c_2, \ldots, c_T)$, each associated with a plot summary s_t and a set of characters $\mathcal{A}_t$. We extract a set of atomic propositions $\mathcal{P} = \{p_1, \ldots, p_N\}$ describing salient narrative facts. For each character $a \in \mathcal{A}$ and proposition $p \in \mathcal{P}$ at time t, we assign an *ignorance state*

$$\sigma_t(a, p) \in \mathcal{S} \times \mathcal{M}, \tag{1}$$

where $\mathcal{S} = \{$UNAWARE, SUSPECTING, KNOWN, MISBELIEVING$\}$ denotes the primary epistemic state and $\mathcal{M} = \{$NONE, WITHHELD, INEXPRESSIBLE$\}$ denotes the optional expression modifier. A character-proposition pair can carry no expression constraint, be deliberately WITHHELD or be INEXPRESSIBLE. These modifiers are mutually exclusive, so a pair cannot fall into both the WITHHELD and INEXPRESSIBLE categories. The collection of all (a, p, σ_t) triples induces a dynamic **Ignorance Graph** $G_t = (V, E_t)$, where vertices $V = \mathcal{A} \cup \mathcal{P}$ and edges $E_t \subseteq \mathcal{A} \times \mathcal{P}$ are labeled with σ_t. The graph evolves across chapters, supporting cross-episode belief continuity.

3.1.1 Proposition Extraction and Normalization

Proposition Extraction and Filtering. We prompt Gemini-3.1-Pro to decompose each plot

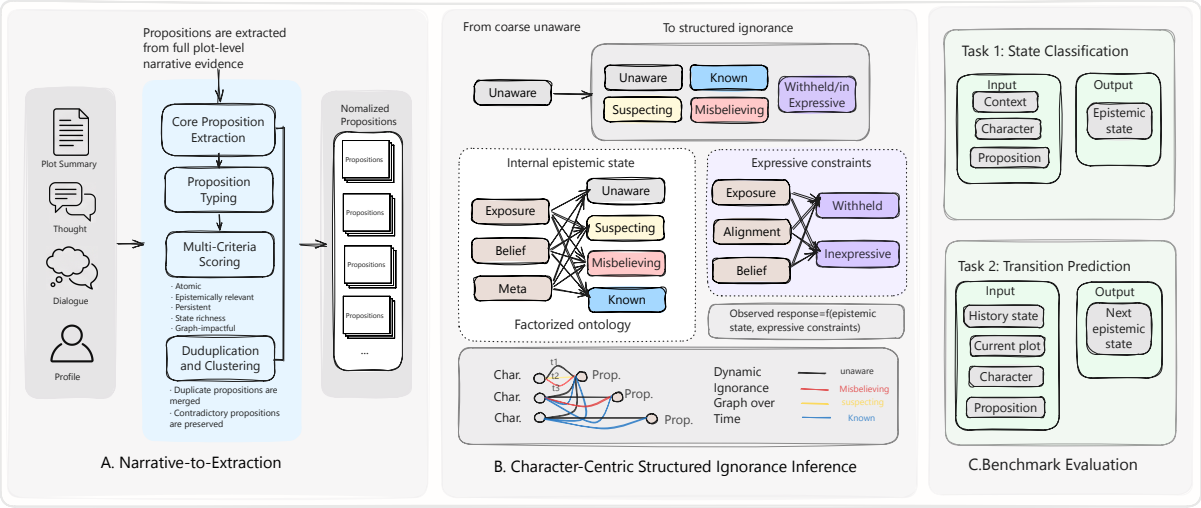


Figure 2: Overview of the IgnoranceBench construction pipeline. Starting from temporally ordered novels, we extract candidate propositions, normalize and deduplicate them, infer character-level ignorance states across chapters, and assemble plot-indexed ignorance graphs for downstream evaluation tasks.

summary s_t into candidate propositions and assess every candidate along five qualitative dimensions, namely *atomicity*, *epistemic relevance*, *temporal persistence*, *state richness*, and *graph-structural impact*, each rated on an ordinal scale $\{0, 1, 2\}$. Only propositions attaining the maximum score on all dimensions are retained, yielding a high-precision set $\mathcal{P}$ in which every p is atomic, epistemically informative, temporally stable, and structurally meaningful for downstream graph construction.

Deduplication. To reduce redundancy while preserving genuine narrative contradictions, we adopt a coarse-to-fine deduplication pipeline, following the common retrieval-then-verification paradigm in semantic matching and textual entailment applications (Reimers and Gurevych, 2019; Dagan et al., 2005; Williams et al., 2018). Candidate pairs are first retrieved using dense embeddings from BGE-M3 (Chen et al., 2024b) with a cosine similarity threshold of 0.80, and then adjudicated by bidirectional natural language inference (NLI) with *roberta-large-mnli*, which builds on RoBERTa pretraining and MNLI-style entailment/contradiction supervision (Zhuang et al., 2021). A pair is merged when both directional entailment probabilities are ≥ 0.85, treating semantic equivalence as approximate bidirectional entailment (Androutsopoulos and Malakasiotis, 2010). Under unidirectional entailment, where one direction is ≥ 0.85 and the reverse is ≤ 0.30, we retain the longer proposition as a more informative surrogate. Pairs whose maximum contradiction probability reaches 0.85 are kept as distinct entries, consistent with NLI-based uses of contradiction detection for factual inconsistency analysis (Kryściński et al., 2020; Laban et al., 2021).

3.1.2 Ignorance State Inference and Graph Construction

We extract propositions from each plot, and conduct character cognition judgment on subsequent propositions taking each of the prior five plots as respective temporal starting points. We annotate each (character a, proposition p, the current plot unit c_t) triple as a point in the state space $\mathcal{S} \times \mathcal{M}$ defined in § 3.1, comprising four mutually exclusive primary states (UNAWARE, SUSPECTING, MISBELIEVING, KNOWN) and one optional, mutually exclusive expression modifier (NONE, WITHHELD, or INEXPRESSIBLE). To make annotation auditable and revisable, we adopt a two-stage pipeline: an LLM annotator produces ordinal scores along interpretable dimensions with grounded textual evidence, and a deterministic procedure projects those scores onto $\mathcal{S} \times \mathcal{M}$. This decoupling confines model variance to the scoring stage and keeps the label assignment fully transparent.

Scoring Dimensions Our schema operationalizes epistemic state attribution along three orthogonal axes together with a reliability term. It separates what a character has access to from what they take to be true, and it also separates epistemic

Task	Book	Samples	Prop.	CPT	Unaw.	Susp.	Misbel.	Known	Withh.	Inexp.
Task 1: Ignorance State Classification	A Thousand Splendid Suns	105	43	552	26	34	19	26	6	8
	Anna Karenina	253	143	1808	63	88	39	63	18	1
	Crime and Punishment	188	64	1040	47	72	22	47	13	1
	Far From the Madding Crowd	265	60	835	66	97	36	66	8	0
	Gone with the Wind	140	96	1221	35	44	26	35	5	0
	Great Expectations	121	81	1197	30	53	8	30	6	2
	Little Women	68	48	780	17	29	5	17	0	0
	Mansfield Park	236	60	881	59	74	44	59	14	1
	Middlemarch	452	133	1826	113	177	49	113	11	4
	North and South	169	63	950	42	52	33	42	8	1
	Pride and Prejudice	285	90	1300	71	71	72	71	12	1
	Sense and Sensibility	193	61	975	48	66	31	48	15	0
	The Age of Innocence	85	49	624	21	34	9	21	10	0
	The Help	100	63	759	25	37	13	25	2	1
	The Kite Runner	32	23	301	8	15	1	8	4	1
	The Song of Achilles	64	41	535	16	18	14	16	2	0
	War and Peace	376	200	3402	94	152	36	94	26	3
	TOTAL	**3132**	**1318**	**18986**	**781**	**1113**	**457**	**781**	**160**	**24**
Task 2: State Transition Prediction	A Thousand Splendid Suns	60	43	552	15	17	13	15	–	–
	Anna Karenina	153	143	1808	38	45	32	38	–	–
	Crime and Punishment	109	64	1040	27	39	16	27	–	–
	Far From the Madding Crowd	173	60	835	43	59	28	43	–	–
	Gone with the Wind	80	96	1221	20	18	22	20	–	–
	Great Expectations	68	81	1197	17	28	6	17	–	–
	Little Women	53	48	780	13	22	5	13	–	–
	Mansfield Park	141	60	881	35	40	31	35	–	–
	Middlemarch	193	133	1826	48	64	33	48	–	–
	North and South	93	63	950	23	28	19	23	–	–
	Pride and Prejudice	200	90	1300	50	47	53	50	–	–
	Sense and Sensibility	129	61	975	32	44	21	32	–	–
	The Age of Innocence	52	49	624	13	18	8	13	–	–
	The Help	48	63	759	12	18	6	12	–	–
	The Kite Runner	20	23	301	5	9	1	5	–	–
	The Song of Achilles	40	41	535	10	11	9	10	–	–
	War and Peace	129	200	3402	32	52	13	32	–	–
	TOTAL	**1741**	**1318**	**18986**	**433**	**559**	**316**	**433**	**–**	**–**

Table 1: Benchmark scale and gold-label counts by novel. Prop. denotes graph propositions and CPT denotes character–proposition–time triples. Task 1 is ignorance state classification and Task 2 is state transition prediction; Unaw., Susp., Misbel., Withh., and Inexp. abbreviate UNAWARE, SUSPECTING, MISBELIEVING, WITHHELD, and INEXPRESSIBLE. Dashes indicate unavailable modifier labels.

content from communicative behavior. This design is consistent with theory-of-mind approaches to narrative understanding (Shinoda et al., 2025) and with communication-theoretic accounts that distinguish strategic nondisclosure from inexpressibility (Grice, 1975; Petronio, 2002). To ground ratings in the source text and reduce hallucination, the annotator additionally extracts typed textual evidence, namely THOUGHT, SPEECH, SUMMARY, or EXPERIENCE, where EXPERIENCE denotes cumulative prior-plot context attributed to the focal character. This evidence is used to substantiate each rating whenever applicable. Six ordinal dimensions are rated on a 0 to 3 scale.

- *Epistemic* variables capture knowledge states. EXPOSURE denotes informational access to P. BELIEF quantifies the doxastic stance toward P across four levels (disbelief, uncommitted, weak belief, strong belief). This signed formulation explicitly represents commitment to $\neg P$ instead of mere belief absence, applying a conservative cap to future events.

- *Communicative* variables model verbalization. EXPRESSION evaluates the explicitness of overt statements about P across four levels (none, vague, clear, full), strictly excluding inner thoughts. ALIGNMENT assesses the coherence between internal thoughts and external speech regarding P across four categories (no evidence, consistency, partial gaps, contradiction).

- *Intentionality* variables distinguish reasons for silence. WITHHOLD quantifies deliberate information concealment despite communicative ability across four levels (none, weak, probable, explicit) to separate strategic silence from incidental omission. INEXPRESSIBLE measures the inability to articulate P due to internal constraints such as trauma or cognitive overload across four similar levels, differentiating genuine inability from unwillingness. These scores are projected to a single final expression modifier, so the two final modifier indicators cannot both be active.

- *Reliability*. EVIDENCE_CONFIDENCE is a global self-report of textual support across all

dimensions.

Detailed rubrics and prompt templates are deferred to Appendix A.

Deterministic Mapping The primary state is determined by the joint configuration of EXPOSURE (e) and BELIEF (b) through monotonic rules that instantiate the gradation of knowledge gaps in narratology (Baboulene, 2017; Abbott, 2013).

- **UNAWARE** applies when $e = 0$, or when $e \geq 1$ and $b = 1$, which indicates suspended judgment. Promotion requires directional commitment.
- **SUSPECTING** and **KNOWN** apply under stronger informational conditions. When exposure is sufficient, namely $e \geq 2$, $b = 2$ yields SUSPECTING and $b = 3$ yields KNOWN. When exposure is partial, namely $e = 1$, the state is capped at SUSPECTING even when $b = 3$, because KNOWN requires both substantial access and firm conviction.
- **MISBELIEVING** applies when $b = 0$ and $e \geq 1$. When $b = 0$ and $e = 0$, the case collapses to UNAWARE, since misbelief presupposes some informational contact that can be misconstrued.

The last rule isolates a distinct doxastic commitment, namely a positive endorsement of $\neg P$. LLM role-players are known to mishandle this pattern in temporally indexed narratives (Ahn et al., 2024), where committed false beliefs are often collapsed into omniscient correctness or conflated with UNAWARE or SUSPECTING.

Modifiers The expression modifier is optional and mutually exclusive. We first assign INEXPRESSIBLE when belief is at least weakly held, so $b \geq 2$, and expression is entirely absent, so *expr* $= 0$, together with a thought–speech gap, so *align* ≥ 2, and a positive inability signal, so *inexpressible* ≥ 1. If this condition is not met, WITHHELD is assigned when belief is firm, so $b = 3$, while expression is suppressed, so *expr* ≤ 1, again with *align* ≥ 2 and a positive intentionality signal, so *withhold* ≥ 2. If neither condition is satisfied, the modifier is NONE. This ordering prevents a single instance from being labeled both WITHHELD and INEXPRESSIBLE; an inability cue takes precedence because WITHHELD presupposes the character can disclose but chooses not to. Requiring both a thought–speech gap and an intentionality cue prevents mere silence from being conflated with deliberate or constrained nondisclosure. The asymmetric belief thresholds, namely $b = 3$ versus $b \geq 2$, reflect the stronger doxastic prerequisite for strategic withholding than for genuine inarticulacy. Full thresholding logic is given in Appendix C.

Temporal Update Given G_{t-1}, we update each edge $(a, p)_t$ using evidence from chapter c_t. A prior belief is carried over only when the current belief is tentative, so $b_t = 2$, and when a high-confidence prior exists, so $b_{t-1} \in \{0, 3\}$ with $\text{EVIDENCE_CONFIDENCE}_{t-1} \geq 2$. In that case, the edge inherits b_{t-1}. EXPOSURE is updated to $\max(e_t, e_{t-1})$. This process yields a temporally consistent sequence $\{G_t\}_{t=1}^{T}$.

3.1.3 Quality Control

We randomly selected the full unsampled ignorance graph of *Pride and Prejudice* for manual annotation verification. We achieved 92.7% accuracy on primary labels and 89.1% on modifiers, and integrated all human-corrected annotations into the official released dataset.

3.2 Benchmark Tasks

IgnoranceBench defines two evaluation tasks over $\{G_t\}$:

Task 1: Ignorance State Classification. Given a character a, proposition p, the current plot unit c_t (including plot summary, character cumulative experiences, and ordered conversations with inner thoughts and dialogues), the model predicts the primary state in $\mathcal{S}$ and the optional expression modifier in $\mathcal{M}$ for $(a, p)_t$.

Task 2: State Transition Prediction. Given $\sigma_{t-1}(a, p)$ and the current plot unit c_t, the model predicts $\sigma_t(a, p)$, evaluating whether models can faithfully track belief evolution.

3.3 Evaluation Metrics

Standard classification metrics (accuracy, recall, macro-F_1) treat all errors symmetrically and cannot diagnose the directional failures characteristic of role-play epistemics. We therefore complement them with a set of role-play-specific metrics, each targeting a specific failure mode identified in Section D.

Source	Model	Acc.	Macro-F_1	Recall	Withheld-F_1	KOAR	IOSR	MKR
Closed-source	GPT-5.4	**68.4**	**64.7**	**64.8**	**81.5**	10.5	**11.0**	7.5
	GPT-4o	51.1	47.3	50.5	60.3	**5.3**	50.1	**5.3**
Open-source	Llama-3.1-70B-Instruct	60.3	52.2	54.3	50.4	6.3	22.2	8.6
	Qwen3-32B	55.6	49.4	53.0	55.7	13.5	31.5	14.5
	Qwen3-14B	50.5	43.7	48.0	30.6	10.3	42.6	12.3
	Gemma-3-12B	48.5	40.5	41.6	64.2	12.5	13.0	18.7
	GLM-4-9B-0414	44.9	37.0	41.4	40.1	9.1	42.9	11.9
	Qwen3-8B	46.8	39.1	43.6	6.7	10.4	41.7	13.5
	Llama-3.1-8B-Instruct	44.3	37.2	39.5	40.5	5.7	34.0	6.7
	Mistral-7B-Instruct-v0.3	35.2	27.0	32.6	8.6	8.7	57.8	13.7
Class-balanced random baseline		27.2	25.0	25.0	9.3	24.9	24.9	24.9

Table 2: Weighted model performance on Task 1 (current-state classification). Accuracy, recall, macro-F_1, and withheld-F_1 are reported as percentages; higher is better. KOAR, IOSR, and MKR are directional error rates; lower is better. Best values are bolded.

Task 1 Ignorance State Classification. We measure three directional error rates and one collapse statistic. **KOAR** (Knowledge Over-Ascription) measures how often non-known cases are leaked as KNOWN, capturing omniscient leakage from the underlying LLM. **IOSR** (Ignorance Over-Suppression) measures how often they are flattened into UNAWARE, capturing the loss of fine-grained ignorance structure. **MKR** (Misbelief-as-Knowledge) characterizes confusion between confident assertion and epistemic correctness.

Task 2 State Transition Prediction. We report **TDA** (Transition Direction Accuracy), the fraction of cases for which the model correctly identifies the direction of epistemic change (improvement, deterioration, or belief correction).

4 Experiments

All reported scores are weighted aggregate percentages from a single evaluation pass over the fixed test set rather than maxima or means over multiple random seeds. Per-novel variation is reported separately in Appendix F.

4.1 Ignorance State Classification

Table 2 details the weighted aggregate results for Task 1. GPT-5.4 achieves the strongest overall performance, with 68.4% accuracy, 64.7 macro-F_1, and 64.8% recall, alongside the highest withheld-modifier F_1 at 81.5. Among open-weight models, Llama-3.1-70B-Instruct achieves the highest primary-state accuracy, whereas Gemma-3-12B remains competitive on the withheld modifier despite weaker main-state predictions. Because the INEXPRESSIBLE modifier has few samples, we exclude its quantitative results from this aggregate analysis and analyze it qualitatively in the case study. Directional metrics show that high accuracy does not prevent epistemic leakage. Leading models still exhibit substantial KOAR and IOSR, indicating persistent knowledge over-ascription and systematic collapse of nuanced ignorance states.

4.2 State Transition Prediction

Table 3 shows that transition prediction remains easier than exact ignorance state classification for the top systems. GPT-5.4 leads by a clear margin with 70.5% accuracy, 68.4 macro-F_1, 69.5% recall, and 83.7 TDA, reliably capturing directions of epistemic change. GPT-4o follows with 64.1% accuracy, 61.8 macro-F_1, and 76.0 TDA. Among open models, Llama-3.1-70B-Instruct performs best with 63.8% accuracy, 60.0 macro-F_1, and 78.2 TDA, ahead of Qwen3-32B at 56.6% accuracy and 52.9 macro-F_1. Other open checkpoints range between 39.4 and 47.8 macro-F_1, with Mistral-7B at the lower bound. Per-novel results are presented in Appendix F.

Source	Model	Acc.	Macro-F_1	Recall	TDA
Closed-source	GPT-5.4	**70.5**	**68.4**	**69.5**	**83.7**
	GPT-4o	64.1	61.8	64.0	76.0
Open-source	Llama-3.1-70B	63.8	60.0	61.6	78.2
	Qwen3-32B	56.6	52.9	56.2	70.6
	Qwen3-14B	52.4	47.8	52.4	67.1
	Gemma-3-12B	48.1	44.7	46.1	71.5
	GLM-4-9B	48.3	45.5	47.7	54.0
	Qwen3-8B	48.4	44.7	47.5	69.9
	Llama-3.1-8B	49.1	44.8	45.9	70.2
	Mistral-7B	43.7	39.4	40.1	63.0
Class-balanced random baseline		26.0	25.0	25.0	–

Table 3: Weighted model performance on Task 2 (state-transition prediction). Accuracy, recall, macro-F_1, and transition direction accuracy (TDA) are reported as percentages.

5 Case Study

The following table provides concrete illustrative analyses for Task 1 and Task 2.

Task 1: Ignorance State Classification	
Case	*Pride and Prejudice*, Plot 7, Catherine Bennet.
Proposition	The youngest Bennet daughters behave foolishly regarding the newly arrived militia regiment.
Context	Mr. Bennet calls his two youngest daughters foolish; Catherine looks down in embarrassment while Lydia remains absorbed in the militia officers.
Key evidence	**Thought:** Catherine fears she may be as silly as Lydia. **Dialogue:** Catherine says she does not know what to say and looks down. **Summary:** Mr. Bennet criticizes his daughters' behavior.
Gold judgment	**State label:** SUSPECTING; **Withheld:** False; **Inexpressible:** True. Catherine suspects the proposition but cannot articulate this self-awareness under embarrassment.
Model outputs	GPT-5.4, Llama-3.1-70B, Qwen3-32B, Qwen3-14B, Gemma-12B, and Llama-3.1-8B predict SUSPECTING; GPT-4o predicts both SUSPECTING and INEXPRESSIVE; GLM-4-9B, Qwen3-8B, and Mistral-7B predict UNAWARE.
Task 2: State-Transition Prediction	
Case	*Pride and Prejudice*, Plot 8, Caroline Bingley.
Proposition	Mr. Darcy is secretly intrigued by Elizabeth Bennet's challenge during their debate about accomplished women.
Context	During the accomplished-women debate, Darcy is inwardly drawn to Elizabeth's wit, while Caroline treats the exchange as a chance to display her own accomplishments.
Key evidence	**Thought:** Caroline aims to underscore Elizabeth's inadequacies. **Dialogue:** Caroline aligns herself with Darcy's high standards. **Summary:** The Bingley sisters criticize Elizabeth behind her back.
Gold judgment	**History state:** UNAWARE; **state label:** MISBELIEVING. Caroline mistakes Darcy's fascination with Elizabeth for dissatisfaction with Elizabeth.
Model outputs	GPT-5.4, GPT-4o, Llama-3.1-70B, Qwen3-32B, and GLM-4-9B predict UNAWARE; Gemma-12B, Llama-3.1-8B, and Mistral-7B predict SUSPECTING; Qwen3-14B and Qwen3-8B predict KNOWN.

Table 4: Compact qualitative examples from *Pride and Prejudice*. The table condenses the annotation context, supporting evidence, gold judgment, and model outputs for Task 1 ignorance state classification and Task 2 state transition prediction.

Ignorance State Classification. The Task 1 column evaluates Catherine Bennet at Plot 7 concerning the proposition that the youngest Bennet daughters behave foolishly toward the militia. Based on Mr. Bennet's criticism alongside Catherine's embarrassed silence and internal doubts, the gold label is SUSPECTING with the INEXPRESSIBLE modifier. This classification indicates that embarrassment hinders clear articulation rather than strategic concealment. Stronger systems including GPT-5.4, GPT-4o, Llama-3.1-70B-Instruct, Qwen3-32B/14B, gemma-3-12B, and Llama-3.1-8B-Instruct successfully recover the main state, with GPT-4o uniquely predicting the inexpressibility modifier. Conversely, GLM-4-9B, Qwen3-8B, and Mistral-7B misclassify the case as UNAWARE, demonstrating local ignorance over-suppression consistent with their high IOSR in Table 2.

State Transition Prediection. The Task 2 column evaluates Caroline Bingley at Plot 8 regarding Mr. Darcy's secret intrigue toward Elizabeth Bennet. The gold state transitions from UNAWARE to MISBELIEVING because Caroline observes their debate but misinterprets Darcy's attention through rivalry to form a committed false belief. No model predicts this exact transition. Despite strong aggregate TDA in Table 3, GPT-5.4, GPT-4o, Llama-3.1-70B-Instruct, Qwen3-32B, and GLM-4-9B exhibit history anchoring by preserving the UNAWARE state. Furthermore, gemma-3-12B-IT, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 output SUSPECTING, capturing uncertainty but missing the false belief. Finally, Qwen3-14B and Qwen3-8B output KNOWN, demonstrating narrator knowledge leakage that aligns with their lower aggregate Task 2 scores.

6 Conclusion

We present the Ignorance Graph, a plot-indexed representation for tracking character-specific ignorance in narrative role-playing. By modeling character-proposition relations over time, the graph separates internal epistemic states from expression-level constraints, allowing evaluation to move beyond whether a character knows a fact and toward whether they are unaware of it, suspect it, misbelieve it, strategically withhold it, or fail to articulate it. Built on this representation, IgnoranceBench introduces two benchmark tasks, ignorance state classification and state transition prediction, over temporally sequenced novels. Experiments with ten large language models show that structured ignorance remains difficult even for the strongest systems. Models can often recover common knowledge states, but they still over-attribute knowledge, suppress fine-grained ignorance into unawareness, and confuse committed false beliefs with either knowledge or uncertainty. These findings suggest that faithful narrative role-playing requires mechanisms that maintain temporally grounded, character-bound epistemic states and communicative constraints, rather than relying only on factual retrieval or surface persona consistency.

Limitations

IgnoranceBench is automatically constructed and therefore inherits noise from proposition extraction, deduplication, and LLM annotation. The INEXPRESSIBLE modifier is also sparse, limiting robust quantitative conclusions. Although the score-to-state projection is deterministic once dimension scores are assigned, both the choice of mapping dimensions and the threshold rules used at state boundaries inevitably involve some subjectivity. We mitigate this issue through explicit rubrics, evidence requirements, and a transparent mapping procedure, but our quality-control verification remains partial. Manual inspection was conducted on the full unsampled ignorance graph of one novel rather than across all books, characters, and state categories. Due to human-resource constraints, we also do not provide human baseline scores for the two benchmark tasks, which limits our ability to separate model difficulty from possible ambiguity in the task formulation.

Future work should further validate the ontology with multiple annotators, report agreement on both dimension scores and final state labels, expand quality checks across the benchmark, and collect human baselines under the same evaluation protocol.

Ethical Considerations

Our study utilizes processed narrative materials from CoSER (Wang et al., 2025a) strictly to construct propositional ignorance graphs and benchmark tasks for academic evaluation. We do not claim ownership of the underlying literary works and strictly prohibit the redistribution of raw novel texts. The artifacts introduced in IgnoranceBench encompass derived annotations, graph structures, and evaluation scripts intended solely to facilitate reproducibility. To ensure compliance with the original access conditions and protect copyright, our public release includes only benchmark metadata and derived labels while excluding source passages. Given that epistemic state annotations could be adapted for unintended purposes, we explicitly restrict the downstream use of IgnoranceBench. Users are prohibited from applying this dataset for impersonating real people, conducting deceptive interactions, invading privacy, or generating harmful simulations.

References

H Porter Abbott. 2013. *Real mysteries: Narrative and the unknowable*. The Ohio State University Press.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

David Baboulene. 2017. Knowledge gaps in popular hollywood cinema storytelling: the role of information disparity. *Education*, 1(1):1.

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.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024a. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. 2023. Propsegment: A large-scale corpus for proposition-level segmentation and entailment recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8874–8893.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024c. Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. Role-playing evaluation for large language models. In *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning*, pages 118–127. Springer.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. 2026. Narrabench: A comprehensive framework for narrative benchmarking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3786–3801.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2016. Tracking the world state with recurrent entity networks. *arXiv preprint arXiv:1612.03969*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan 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*, pages 14397–14413.

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.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. *Preprint*, arXiv:2308.07702.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.

Emre Kuru, Anıl Doğru, Merve Doğan, and Reyhan Aydoğan. 2025. Evaluating theory-of-mind in large language models through opponent modeling. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, pages 1–9.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Preprint*, arXiv:2111.09525.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8211–8225.

Lingyu Li, Yixu Wang, Haiquan Zhao, Shuqi Kong, Yan Teng, Chunbo Li, and Yingchun Wang. 2024. Reflection-bench: evaluating epistemic agency in large language models. *arXiv preprint arXiv:2410.16270*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. In *International Conference on Learning Representations*, volume 2024, pages 1892–1915.

Sandra Petronio. 2002. *Boundaries of privacy: Dialectics of disclosure*. Suny Press.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 8(9).

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models ?lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1520–1528.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, and 1 others. 2025a. Coser: Coordinating llm-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1840–1873.

Ye Wang, Tong Li, Meixuan Li, Ziyue Cheng, Ge Wang, Hanyue Kang, Yaling Deng, Hongjiang Xiao, and Yuan Zhang. 2025b. Rvbench: Role values benchmark for role-playing llms. *Computers in Human Behavior: Artificial Humans*, page 100184.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models ?theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

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*, pages 10691–10706.

Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, and Song Liu. 2025. Dmt-rolebench: A dynamic multi-turn dialogue based benchmark for role-playing evaluation of large language model and agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25760–25768.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, and 1 others. 2024. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, and 1 others. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, pages 1218–1227.

A Annotation Rubric and Dimension Definitions

This appendix reports the annotation rubric and the explicit dimension definitions used in prompting. The rubric is shared by human annotation and automatic scoring so that the evidence standards are aligned. All dimensions are scored on a four-level ordinal scale from zero to three.

Dimension definitions used in scoring prompts.

- EXPOSURE: degree to which a character is exposed to proposition-level information in the current plot unit.

- BELIEF: strength of commitment that the proposition is true.
- EXPRESSION: explicitness of outward verbalization of the proposition.
- ALIGNMENT: consistency between internal thought evidence and spoken evidence for the same proposition.
- WITHHOLD: evidence of deliberate nondisclosure when the character can reasonably disclose.
- INEXPRESSIBLE: evidence that articulation is infeasible for the character, independent of willingness.

The final expression modifier is projected from the WITHHOLD and INEXPRESSIBLE scores and is constrained to one of NONE, WITHHELD, or INEXPRESSIBLE.

B Prompt Templates

This appendix reproduces the prompt templates and lists the task constraints used in our annotation pipeline.

Proposition Extraction Prompt

Task: Extract 1–3 atomic, canon-true propositions for tracking character ignorance states and expression constraints (*unaware, suspecting, misbelieving, known, withheld, inexpressible*). If the text lacks epistemic value (e.g., pure atmosphere), output "-".
Input: `{plot_summary}`, `{char_part}`, `{profile_part}`, `{conv_part}`
Core Constraints:

- **Grounded & Canon-True:** Propositions must be entailed by the summary. Use profiles to verify book-level truth. Rewrite false surface beliefs/rumors into underlying canon-true facts.
- **Local Verifiability:** Must be verifiable using only the provided plot text.
- **Atomicity:** One sentence = ONE main claim (simple present tense, resolved referents). Strip conversational framing (e.g., extract the hidden fact Z, not "A tells B about Z").
- **Plot-Advancing:** Must be a plot-driving hinge (decisions, hidden agendas). Prioritize secret motives over surface actions.
- **Categorization:** Assign one type: `plot_event_anchor`, `epistemic_shift`, or `concealment_misbelief_suspicion_trigger`.

Scoring (0–2 scale): Evaluate each candidate on `ATOMICITY`, `EPISTEMIC_RELEVANCE`, `TEMPORAL_PERSISTENCE`, `STATE_RICHNESS`, and `GRAPH-STRUCTURAL_IMPACT`. Prefer candidates with score 2 across dimensions.

Output Format (Strict):
For $i = 1..K$ (where $K \in [1, 3]$):

```
SELECTED_PROPOSITION_i: <one declarative sentence>
SELECTED_TYPE_i: <type>
ATOMICITY_i: <0, 1, or 2>
EPISTEMIC_RELEVANCE_i: <0, 1, or 2>
TEMPORAL_PERSISTENCE_i: <0, 1, or 2>
STATE_RICHNESS_i: <0, 1, or 2>
GRAPH-STRUCTURAL_IMPACT_i: <0, 1, or 2>
```

Ignorance-Scoring Prompt

Role. You are an expert at analyzing character knowledge and ignorance in narrative fiction.
Input. A narrative unit consisting of:

- a plot summary (objective narration);
- cumulative past experiences of each character appearing in this plot;
- the plot's conversations in order, each with scene, topic, inner THOUGHT, and spoken SPEECH.

Conventions.

- `[...]` denotes inner thought.
- `(...)` attached to dialogue denotes manner or stage direction.
- Both are valid evidence.

Target proposition P. `{prop.text}`
Characters to score. `{char_list}` Ignore all other characters.
Task. For each listed character, adopt that character's in-plot perspective and assign integer scores with textual evidence for the following dimensions.

1. **Exposure (0–3):** informational access to P (heard, read, saw, or was told). Exposure is not belief. Exclude narrator-only facts unless the character plausibly learns them.
   - *0*: none
   - *1*: vague or peripheral hints
   - *2*: clear but non-central access
   - *3*: first-hand access or explicit statement of P
2. **Belief_p (0–3):** doxastic stance toward P (truth, not desire or approval). For propositions about another agent's mental state, score whether the character believes that state truly holds. There is no neutral midpoint; under weak evidence, prefer 1.
   - *Pitfalls:* intention is not expectation; wish is not belief; for not-yet-occurred future events, cap at 2 unless objectively settled (e.g., signed, irreversible, inevitable) or retrospectively narrated.
   - *0*: disbelieves
   - *1*: uncommitted or suspended
   - *2*: weakly believes (guess, suspicion, "maybe")
   - *3*: strongly believes (firm, confirmed)

3. **Expression (0–3):** outward speech conveying P. THOUGHT alone does not count; bare manner cues without propositional content do not count.
   - *0*: none
   - *1*: indirect or euphemistic expression
   - *2*: substance understandable
   - *3*: explicit statement or clear paraphrase
4. **Alignment (0–3):** relation between thought and speech about P. Thought-only usually implies a partial gap; speech without inner evidence implies 0.
   - *0*: no inner evidence
   - *1*: consistent
   - *2*: partial gap (speech understates, evades, or thought adds detail)
   - *3*: contradiction
5. **Withhold (0–3):** deliberate concealment of P from someone who would care. Mere privacy or not being asked does not count. Withhold means *will not tell*.
   - *0*: none
   - *1*: light avoidance or deflection
   - *2*: probable intentional nondisclosure
   - *3*: explicit choice not to reveal
6. **Inexpressible (0–3):** inability to articulate P (e.g., overwhelm, trauma, taboo, conceptual limits). This is distinct from unwillingness. Inexpressible means *cannot*.
   - *0*: none
   - *1*: struggle cues
   - *2*: failed attempts or breakdown
   - *3*: stated inability
7. **Evidence_confidence (0–3):** overall textual reliability across all dimensions.
   - *0*: none
   - *1*: weak
   - *2*: moderate
   - *3*: strong

Evidence requirements.

- Only consider characters listed above; for any other names, ignore or treat as noise.
- For each scored dimension, you MUST provide at least one short evidence string that justifies the score.
- Include all strongest relevant snippets, not just one.
- Each evidence string must begin with one of the following typed prefixes:
  - `THOUGHT - <Name>: ...`
  - `SPEECH - <Name>: ...`
  - `SUMMARY: ...`
  - `EXPERIENCE: ...`
- ...

Output format. Output a single valid JSON array only, with no comments, no trailing commas, and no extra text. Include one object per character and follow this schema exactly:
`{example}`
Context. `{context}`
After the marker `{OUTPUT_SECTION_MARKER}`, output only the JSON array.

C Score-to-State Mapping

Algorithm 1 presents the deterministic projection from rubric scores to the primary state and optional expression modifier.

Algorithm 1 Score-to-State Mapping

Require: Tuple $(e, b, x, a, w, n) \in \mathbb{Z}^6_{\geq 0}$:
e: EXPOSURE, b: BELIEF, x: EXPRESSION,
a: ALIGNMENT, w: WITHHELD, n: INEXPRESSIBLE
Ensure: Primary state $s \in \mathcal{S} \triangleq \{\text{UNAWARE}, \text{SUSPECTING}, \text{MISBELIEVING}, \text{KNOWN}\}$; modifier indicators $(m_w, m_n) \in \{(0,0), (1,0), (0,1)\}$

// Stage 1: Determine primary state from (e, b)
Initialize $s \leftarrow$ UNAWARE
if $e \geq 1$ **then**

$$s \leftarrow \begin{cases} \text{MISBELIEVING}, & b = 0 \\ \text{UNAWARE}, & b = 1 \\ \text{SUSPECTING}, & b = 2 \ \vee \ (e = 1 \wedge b \geq 2) \\ \text{KNOWN}, & e \geq 2 \wedge b \geq 3 \end{cases}$$

end if

// Stage 2: Compute mutually exclusive modifier
$m_w \leftarrow 0;\ m_n \leftarrow 0$
if $b \geq 2 \wedge x = 0 \wedge a \geq 2 \wedge n \geq 1$ **then**
  $m_n \leftarrow 1$
else if $b = 3 \wedge x \leq 1 \wedge a \geq 2 \wedge w \geq 2$ **then**
  $m_w \leftarrow 1$
end if
return $(s,\ m_w,\ m_n)$

D Evaluation Metrics

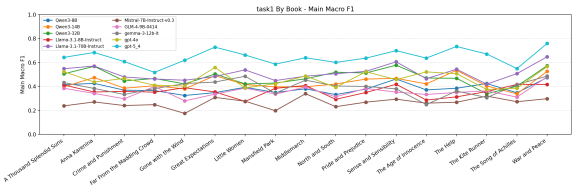
Let $\mathcal{Y} = \{U, S, M, K\}$ denote the main states. Here U is unaware, S is suspecting, M is misbelieving, and K is known. Let $C(y, \hat{y})$ be the confusion matrix.

Task 1 diagnostics. KOAR measures over-prediction of known from non-known states:

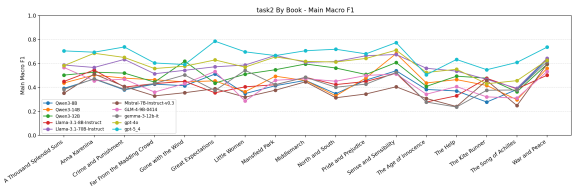
$$\text{KOAR} = \frac{\sum_{y \in \{U,S,M\}} C(y, K)}{\sum_{y \in \{U,S,M\}} \sum_{\hat{y} \in \mathcal{Y}} C(y, \hat{y})}.$$

IOSR measures flattening to unaware from non-unaware states:

$$\text{IOSR} = \frac{\sum_{y \in \{S,M,K\}} C(y, U)}{\sum_{y \in \{S,M,K\}} \sum_{\hat{y} \in \mathcal{Y}} C(y, \hat{y})}.$$



(a) Task 1 macro-F_1 by novel.



(b) Task 2 macro-F_1 by novel.

Figure 3: These line plots compare macro-F_1 scores across novels for multiple models in Task 1 and Task 2, showing performance variability by book and consistent relative rankings across tasks.

MKR measures promotion of misbelieving to known:

$$\mathrm{MKR} = \frac{C(M,K)}{\sum_{\hat{y}\in\mathcal{Y}} C(M,\hat{y})}.$$

Task 2 diagnostic. Let p be the previous state, y the gold next state, and $\hat{y}$ the predicted next state. We use ranks $r(U) = 0, r(S) = 1, r(M) = 1$, and $r(K) = 2$. **TDA** is transition direction accuracy:

$$\mathrm{TDA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\Big[d(y^{(i)}, p^{(i)}) = d(\hat{y}^{(i)}, p^{(i)})\Big],$$

$$d(a,b) = \mathrm{sgn}(r(a) - r(b)).$$

E Additional Label Distribution and Model Performance Variation Across Novels

Figure 3 reports the corresponding per-novel macro-F_1 trajectories. The paired line plots illustrate model performance on Task 1 and Task 2 measured by macro-F_1 across novels. While the same relative model performance hierarchy holds across both tasks, with GPT-5.4 consistently leading, scores vary notably by novel. *War and Peace* yields high performance across models in both tasks, while other novels present greater challenges. The patterns of performance fluctuation across novels are largely consistent between Task 1 and Task 2.

Figure 4 supplements the aggregate label counts in Table 1 with dataset-level density views. The two heatmaps illustrate the per-book distribution of six belief-state label proportions across 17 novels in Task 1 and Task 2, highlighting notable differences in label prevalence, particularly for the withheld and inexpressible categories between the two tasks.

F Additional Novel-Level Results

Per-novel performance varies substantially, reflecting differences in narrative density, character set size, the availability of explicit mental-state evidence, and the unequal volume of annotated data generated from different novels. Table 5 summarizes the highest-scoring Task 1 model for each novel. For Task 1, GPT-5.4 is the strongest model on every novel, although its absolute scores still vary widely across narrative settings. This compact table captures the main cross-novel pattern at a glance. The same trend holds: GPT-5.4 stays on top, while the hardest novels remain the ones with the sparsest or most deceptive evidence.

Table 6 reports the highest-scoring Task 2 model for each novel. GPT-5.4 achieves the best macro-F_1 on fourteen of seventeen novels, while Qwen3-32B leads on *Gone with the Wind* and Llama-3.1-70B-Instruct leads on *Mansfield Park* and *The Age of Innocence*. These exceptions suggest that open-weight models can partially recover when the problem is framed as explicit state-transition tracking rather than direct state reconstruction.

Novel	Best Model	Acc.	Macro-F_1	Recall	TDA
A Thousand Splendid Suns	GPT-5.4	71.7	70.5	71.6	90.0
Anna Karenina	GPT-5.4	71.9	69.4	70.8	89.5
Crime and Punishment	GPT-5.4	74.3	73.8	73.6	83.5
Far From the Madding Crowd	GPT-5.4	65.3	60.4	62.4	75.7
Gone with the Wind	Qwen3-32B	66.2	61.9	65.6	73.8
Great Expectations	GPT-5.4	**77.9**	**78.5**	**78.9**	86.8
Little Women	GPT-5.4	73.6	69.7	69.3	90.6
Mansfield Park	Llama-3.1-70B-Instruct	66.7	66.7	66.8	73.8
Middlemarch	GPT-5.4	70.5	70.6	71.2	84.5
North and South	GPT-5.4	73.1	71.9	72.1	83.9
Pride and Prejudice	GPT-5.4	69.5	68.0	69.4	83.0
Sense and Sensibility	GPT-5.4	77.5	77.4	76.4	84.5
The Age of Innocence	Llama-3.1-70B-Instruct	67.3	56.0	60.4	78.8
The Help	GPT-5.4	70.8	63.3	65.3	87.5
The Kite Runner	GPT-5.4	70.0	54.8	58.9	**95.0**
The Song of Achilles	GPT-5.4	65.0	60.9	64.4	90.0
War and Peace	GPT-5.4	73.6	73.6	74.3	89.1

Table 6: Best per-novel Task 2 results. Rows select the highest macro-F_1 model; all scores are percentages.

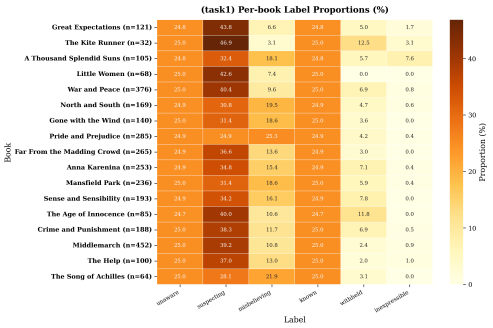
G Confusion Analysis

Figure 5 illustrates classification performance across four epistemic states. The diagonal entries show that most systems identify the UNAWARE state with relatively high accuracy. GPT-5.4 has the strongest overall performance, maintaining high true-positive rates for UNAWARE, SUSPECT-

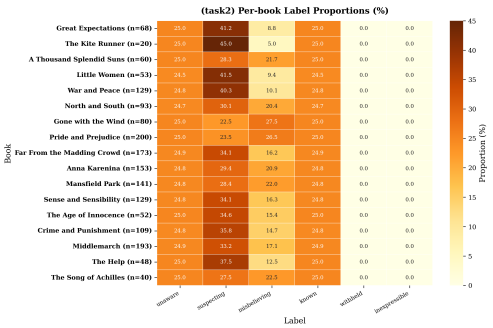
Novel	Best Model	Acc.	Macro-F_1	Withheld-F_1	KOAR	IOSR	MKR
A Thousand Splendid Suns	GPT-5.4	65.7	64.3	44.4	13.9	7.6	10.5
Anna Karenina	GPT-5.4	73.1	68.3	85.7	9.5	4.2	12.8
Crime and Punishment	GPT-5.4	69.2	60.7	85.7	12.1	11.3	9.1
Far From the Madding Crowd	GPT-5.4	61.1	51.7	71.4	14.1	20.6	27.8
Gone with the Wind	GPT-5.4	64.3	61.9	80.0	14.3	14.3	7.7
Great Expectations	GPT-5.4	73.6	72.8	90.9	17.6	9.9	**0.0**
Little Women	GPT-5.4	**83.8**	66.2	0.0	7.8	**2.0**	**0.0**
Mansfield Park	GPT-5.4	57.6	58.6	75.0	13.0	10.7	4.6
Middlemarch	GPT-5.4	68.1	63.9	90.0	10.3	12.4	6.1
North and South	GPT-5.4	63.3	60.1	87.5	11.8	13.4	12.1
Pride and Prejudice	GPT-5.4	64.2	63.6	80.0	8.9	14.9	2.8
Sense and Sensibility	GPT-5.4	71.5	69.8	84.6	**1.4**	11.7	**0.0**
The Age of Innocence	GPT-5.4	71.8	63.6	81.8	4.7	6.2	**0.0**
The Help	GPT-5.4	77.0	73.3	**100.0**	12.0	2.7	15.4
The Kite Runner	GPT-5.4	56.2	67.0	85.7	33.3	12.5	**0.0**
The Song of Achilles	GPT-5.4	64.1	54.9	0.0	12.5	10.4	14.3
War and Peace	GPT-5.4	77.7	**75.8**	87.5	6.0	7.8	**0.0**

Table 5: Best per-novel Task 1 results. Rows report the model with the highest macro-F_1 for each novel. Accuracy, macro-F_1, and withheld-F_1 are percentages; KOAR, IOSR, and MKR are directional error rates.

ING, and KNOWN. In contrast, MISBELIEVING remains difficult for all evaluated systems: true MISBELIEVING instances are often misclassified as either SUSPECTING or UNAWARE. Smaller open-weight models such as Mistral-7B and Llama-8B further degrade on the boundary between UNAWARE and SUSPECTING, often defaulting to UNAWARE. This pattern suggests that language models can recognize basic knowledge absence more reliably than nuanced cognitive states involving incorrect beliefs.

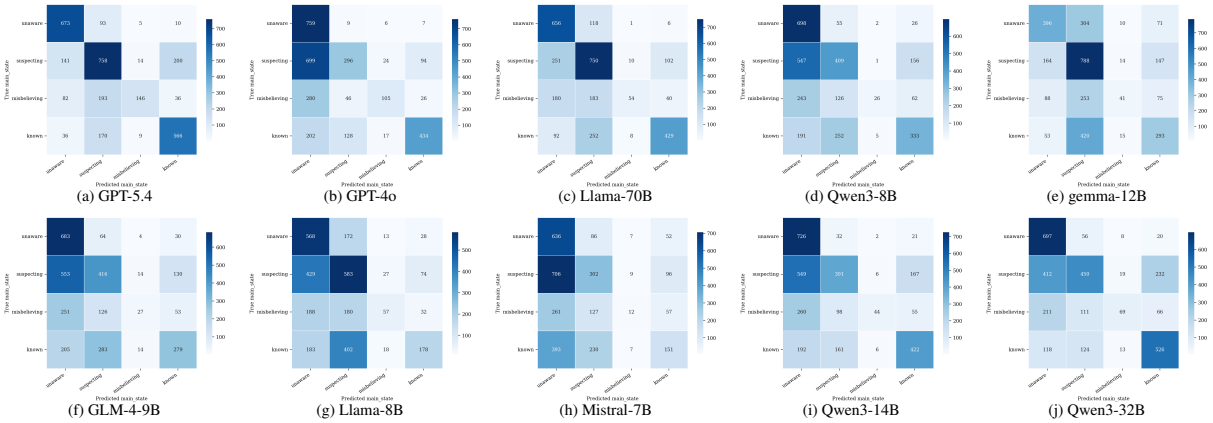


(a) Task 1 label distribution proportions by novel.



(b) Task 2 label distribution proportions by novel.

Figure 4: Heatmaps show per-book label proportions (%) for the six belief states across novels in Task 1 and Task 2, revealing distinct label distribution patterns between the two tasks.



(a) GPT-5.4 (b) GPT-4o (c) Llama-70B (d) Qwen3-8B (e) gemma-12B

(f) GLM-4-9B (g) Llama-8B (h) Mistral-7B (i) Qwen3-14B (j) Qwen3-32B

Figure 5: Overall Task 1 confusion matrices for closed and open-weight systems. Matrices are row-normalized to show the distribution of predicted mental states given each true mental state class. Diagonal entries represent correct predictions, while off-diagonal entries indicate misclassifications. Color intensity corresponds to count magnitude, with darker shades denoting higher frequencies.