

# Beyond Math: Stories as a Testbed for Memorization-Constrained Reasoning in LLMs

Yuxuan Jiang , Francis Ferraro

Department of Computer Science and Electrical Engineering  
University of Maryland, Baltimore County  
yuxuanj1@umbc.edu

## Abstract

Memorization has been shown to greatly inflate Large Language Models’ (LLMs) performance on domains such as math and logic, where success should primarily rely on applying generalizable reasoning rules. In many real-world applications, however, memorization is not meant to be eliminated but *selectively constrained*—for example, in story understanding, where background knowledge must be integrated with narrative context. Drawing on the cognitive science distinction between “verbatim” (exact recall) and “gist” (semantic abstraction) memorization, we propose a two-tier framework for analyzing how LLMs reason under different degrees of memory access. The **Inductive (prompt-guided) Setting** softly steers models to reason through selective, context-relevant recall, while the **Restrictive Setting** imposes stronger constraints by limiting verbatim memory access. Evaluating GPT-4o, LLaMA3.3-70B, and DeepSeek V3 on six character-centric story understanding benchmarks, we find up to a 45.2% accuracy drop under the Restrictive Setting, revealing strong dependence on surface recall. By contrast, the Inductive Setting maintains performance, indicating that prompting can align LLMs toward memorization-constrained reasoning.

## 1 Introduction

Large Language Models (LLMs) have achieved remarkable success on a wide range of reasoning benchmarks. However, the manner in which these successes are achieved is often less satisfying. This can be especially salient in character-centric story understanding tasks, one of which we show in Figure 1: there, we see how a model can correctly solve the task of identifying a character in a dialogue. However, it does so by simply recalling the original script verbatim—a form of *memorization* that deviates from the intended reasoning behavior. As we see, once the character names are replaced,

this shortcut collapses, forcing the model to rely on contextual cues such as a “dramatic and playful” tone to recover the correct speaker through genuine reasoning. This transition under memorization constraints raises a deeper question: *Can LLMs reason in the way humans expect—integrating selective background knowledge with narrative context rather than relying solely on surface recall?*

Character-centric story understanding provides an ideal testbed for probing this issue. Such tasks demand combining background knowledge with story context to infer characters’ motivations, emotions, and relationships (Yu et al., 2023a; Vallurupalli et al., 2024; Shen et al., 2024; Brahman et al., 2021). Because many benchmarks originate from well-known media sources, LLMs may have encountered them verbatim during pretraining (Chang et al., 2023a). When characters’ names or stylistic cues are altered, memorization shortcuts collapse, exposing the model’s true reasoning capacity.

Prior work has examined the tension between memorization and reasoning mainly in domains such as math and logic, where models are expected to apply general rules and treat near-duplicates as new problems (Lesci et al., 2024; Jin et al., 2024; Carlini et al., 2023; Prabhakar et al., 2024; Chen et al., 2024). In contrast, many real-world applications require what we call *memorization-constrained reasoning*, where memorization is not entirely undesirable but must be used *selectively*—retrieving bounded, context-relevant knowledge and integrating it with local evidence. The challenge, therefore, is not to eliminate memory, but to understand how reasoning operates when memorization is constrained rather than removed.

To examine this systematically, we propose a **graded intervention framework** inspired by the cognitive distinction between *verbatim* (exact recall) and *gist* (semantic abstraction) memory. The **Inductive Setting** preserves the original text but introduces *gist-based prompting* to encourage selec-

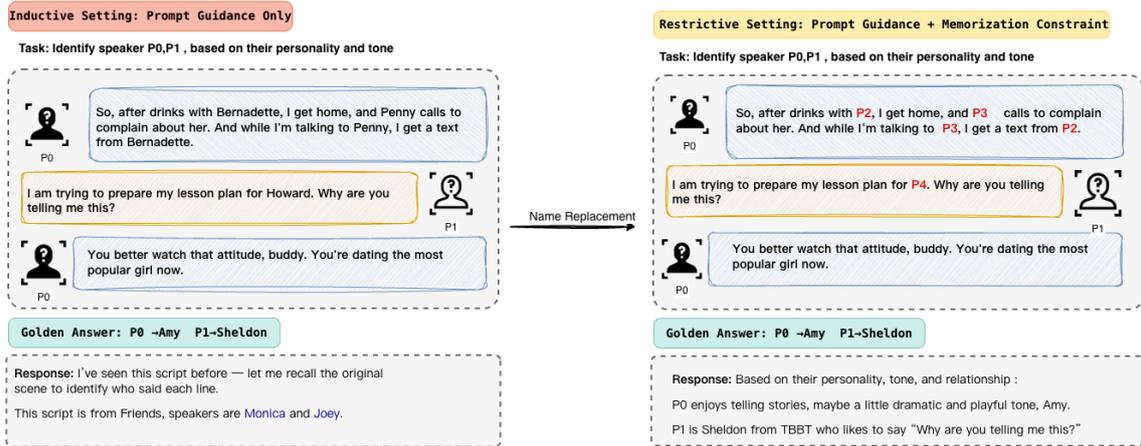


Figure 1: Illustration of our graded intervention framework. The **Inductive Setting** encourages selective, context-based reasoning, while the **Restrictive Setting** limits verbatim recall by replacing character names.

tive, context-driven recall—probing whether reasoning can be *elicited* through guided induction. The **Restrictive Setting** imposes a stronger constraint: all character names are replaced with culturally distinct alternatives, suppressing direct memory access to test whether reasoning can be *sustained* when memorization is blocked. Together, these two settings form a continuum of interventions that distinguish between *behavioral steerability* (whether models can be guided to reason as intended) and *intrinsic robustness* (whether reasoning endures under strict memory constraints).

Applying this framework to six character understanding tasks with three representative LLMs (GPT-4o, LLaMA3-70B, and DeepSeek V3), we find clear support for our hypothesis. Performance drops by as much as 45.8% in the **Restrictive Setting**, showing that current benchmarks substantially overestimate reasoning by allowing verbatim memorization to dominate. By contrast, accuracy in the **Inductive Setting** remains largely stable, indicating that prompting can guide models toward reasoning that appropriately integrates memorization rather than relying on surface recall. Together, these findings reveal not only a critical gap in how current evaluations interpret “reasoning,” but also a concrete way to probe and improve it.

In summary, our key contributions are:

1. **Expanding from Reasoning to Understanding Tasks:** We extend the study of memorization and reasoning beyond math and logic to *character-centric story understanding*, a domain that requires controlled use of memorization for reasoning.

2. **An Inductive–Restrictive Framework:** We propose a two-tier design: the **Inductive Setting** guides models toward memorization-constrained reasoning, while the **Restrictive Setting** limits verbatim recall to test intrinsic reasoning ability.

3. **Diagnosing and Steering LLM Reasoning:** Our results reveal that restricting verbatim cues causes large performance drops, while inductive prompting maintains accuracy—showing how prompting can steer models toward the intended, constrained form of reasoning.

## 2 Related Work

### 2.1 Memorization and Reasoning in LLMs

LLMs’ impressive reasoning capabilities have attracted significant attention (Tong et al., 2023; Yang et al., 2023b; Xiong et al., 2024; Tan et al., 2024; Jiang et al., 2025; Jiang and Ferraro, 2026; Xu et al., 2025; Zhang et al., 2025), yet their behavior is also strongly shaped by memorization (Zhou et al., 2023; Chen et al., 2024; Biderman et al., 2024). Prior studies show that memorizing training data can inflate benchmark scores (Khandelwal et al., 2020; Blevins and Zettlemoyer, 2022) and that conflicting memories can even degrade performance (Su et al., 2024; Wang et al., 2024; Yang et al., 2025).

Cognitive science offers a useful lens for this distinction, separating *verbatim memorization*—precise recall of surface details—from *gist memorization*—semantic abstraction of meaning (Reyna and Brainerd, 1998; Brainerd and

Reyna, 2002). To avoid confusion with architectural notions of memory in NLP (Alizadeh et al., 2024; Kwon et al., 2023), we use these terms to describe distinct *recall styles* rather than neural mechanisms. This distinction underpins our analysis of how LLMs balance shallow recall with deeper reasoning.

Recent NLP research has sought to disentangle and quantify memorization (Lesci et al., 2024; Jin et al., 2024; Carlini et al., 2023; Prabhakar et al., 2024), examining phenomena such as multifaceted memorization (Prashanth et al., 2024; Li et al., 2024) and logit-based tracing (Chen et al., 2024). While mitigation methods like out-of-distribution filtering (Biderman et al., 2024) help in structured reasoning tasks, they are less applicable to character-centric domains grounded in known fictional texts. In contrast, we propose a prompting-based framework that *tests whether LLMs can regulate their use of memorization under explicit constraints*, distinguishing verbatim recall from reasoning that integrates bounded, context-relevant memory.

## 2.2 Character centric Story Understanding and Its Evaluation with LLMs

Character understanding has long been a focus in psychological, literary, and educational research (Bower and Morrow, 1990; McKee, 2010; Paris and Paris, 2003). It spans tasks such as character recognition, coreference resolution, summarization, role detection, goal modeling, personality analysis, and question answering (Bamman et al., 2020; Chen et al., 2022; Frermann et al., 2018; Vallurupalli et al., 2024; Yu et al., 2023b).

Recent work demonstrates that LLMs achieve strong results on character-centric benchmarks (Li et al., 2023; Stambach et al., 2022; Goyal et al., 2023). However, many of these datasets are derived from publicly available or widely consumed fictional sources, such as movie scripts, TV shows, and novels (Sang et al., 2022a; Tapaswi et al., 2016; Caselli and Vossen, 2017), which makes them vulnerable to pretraining overlap and surface-level memorization (Chang et al., 2023a). Although prior research has studied LLM memorization broadly (Carlini et al., 2023; Xie et al., 2024; Yang et al., 2023a), little is known about how it specifically affects character understanding tasks. Our work fills this gap by examining memorization’s role in these tasks and proposing interventions to distinguish between recall and reasoning.

## 3 Data and Experiment Setup

To comprehensively evaluate character understanding in LLMs, we compile two datasets: **CharBench** and **CharScript**. Inspired by NLP benchmark collections such as GLUE (Wang et al., 2018), our **CharBench** integrates six publicly available datasets spanning character-centric tasks like coreference resolution, personality inference, and dialogue-based QA (details in Section 6.1).

The second dataset, **CharScript**, is newly constructed for this study by systematically sampling textual segments from original fictional sources underlying CharBench, explicitly probing models’ verbatim memorization (details in Section 3.2).

### 3.1 CharBench: A Benchmark Collection for Character Understanding

CharBench aggregates six widely recognized NLP benchmarks covering diverse narrative mediums—TV shows, movies, and novels. To ensure comparability, we retain each dataset’s original splits, preprocessing, and evaluation metrics. Detailed task descriptions and implementations are provided in Appendix E.

**Benchmark Sources** Following established NLP benchmark practices (Wang et al., 2018), CharBench includes:

- **MovieCoref** (Chen and Choi, 2016): Coreference resolution in movie scripts.
- **TVSHOWGUESS** (Sang et al., 2022b): Speaker identification in TV dialogues.
- **PERSONET** (Yu et al., 2023b): Personality inference from literary passages.
- **CSI Corpus** (Frermann et al., 2018): Role detection from crime shows.
- **FriendsQA** (Yang and Choi, 2019): Multi-turn dialogue-based QA.
- **SummScreen** (Chen et al., 2022): TV episode summarization.

**Evaluation Metrics** We strictly follow original dataset metrics (exact match, ROUGE, F1 score), as detailed in Table 1.

### 3.2 CharScript: Fictional Works Memorization Benchmark Dataset

Motivated by recent NLP research on memorization in language models (Chang et al., 2023b), we constructed the CharScript dataset to evaluate whether LLMs can recognize the source of fictional text segments. Following a similar setup to (Chang

Task Name	Task Type	Data Source	Metric
TVShow Guess	Character Guessing	TV Shows	Exact Match
FriendsQA	Question Answering	TV Shows	Exact Match
Screenplay	Character Coreference	Movies	F1 Score
SummScreen	Text Summarization	TV Shows	ROUGE
CSI	Role Detection	TV Shows	F1 Score
PERSONET	Personality Understanding	Novels	Exact Match

Table 1: Overview of benchmark tasks, including their type, data sources, and evaluation metrics.

et al., 2023b), we probe whether models can correctly attribute short passages to their original narrative sources, as an indicator of verbatim memorization. We present our experimental design and findings in Section 5.1.

We systematically sampled and extracted **200 textual segments**, averaging approximately **312 words each**, from **40 diverse fictional sources** (11 TV series, 9 movies, and 20 novels). These sources were carefully selected to maximize narrative diversity, textual complexity, and representativeness of popular fictional works included in CharBench. Each source contributes exactly **5 segments**, structured uniformly in the format *[Title]; Scene context*, combining both original dialogue and narrative descriptions.

### 3.3 Models

We evaluate three frontier LLMs—open-source LLaMA 3.3-70B (AI, 2024) and DeepSeek V3 (Liu et al., 2024), as well as the closed-source GPT-4o (OpenAI, 2023)—via API-based inference. To address stochastic variability, each model was run three times per setting, and average performance is reported. The total cost of our experiments was approximately \$100.

## 4 Inductive Setting: Guiding Memorization-Constrained Reasoning

In this setting, we aim to examine whether LLMs can follow explicit prompt instructions to regulate their use of memorization and reasoning as intended. Specifically, we test whether prompting alone—without altering the input text—can induce models to move away from verbatim recall and instead engage in reasoning that relies on relevant, context-bound memory and semantic abstraction.

### 4.1 Character Guessing Task Format

We select the Character Guessing Task specifically because it closely mirrors the name cloze tests (Chang et al., 2023b) previously established as effective probes for verbatim memorization in

LLMs. Name cloze tests measure memorization by removing character names from familiar contexts and testing if models can accurately recall them. Similarly, in our task, speaker labels are removed from dialogue excerpts, requiring the model to infer speakers solely from textual context. If models rely heavily on verbatim memorization, accuracy should significantly degrade when surface cues—such as explicit character names—are disrupted. Conversely, if models genuinely reason about character interactions and relationships, performance should remain relatively stable despite these perturbations.

Following the original TVShow Guess setup (Sang et al., 2022c), the model is tasked with predicting the speaker of each utterance in a dialogue segment where all speaker labels are removed.

### 4.2 Prompting Strategies

We design two distinct prompting conditions to examine how LLMs rely on different memorization mechanisms when identifying characters. Verbatim prompting encourages mechanical memorization, where the model recalls exact substrings from its training data and reconstructs missing parts without reasoning. In contrast, gist prompting focuses on extracting key information and performing memorization-based matching. For character understanding tasks, the most relevant gist cues typically include character relationships, significant events, and personality traits.

All prompts share the same base instruction: *"The following is a dialogue. Please identify who they are. . . ."*

The two variations are as follows:

**Verbatim Prompt:** *"...Do this by direct memorization retrieval—do not infer based on traits, interactions, or context. Recall names exactly from past dialogue knowledge without explanations or reasoning."*

**Gist Prompt:** *"...Do this by analyzing relationships, key events, and personality traits to match them with known characters. Do not directly retrieve from memorization who said the sentences."*

**Prompt Robustness** To positively establish that our findings are stable across wording choices, we designed two paraphrastic variants for each condition (*verbatim* and *gist*), keeping the instructional structure fixed while slightly altering phrasing (Appendix H). For each model, we ran three trials

per variant and report mean accuracy. Across all models, the qualitative ordering observed in the main results is consistently reproduced under these nearby formulations. Numerical results appear in Appendix Table 9.

### 4.3 Result and Analysis

Our key observations are as follows:

- **Baseline Performance:** The baseline condition reflects the overall performance of the models, with GPT-4o demonstrating a clear lead to others, while Llama 3.3-70B and DeepSeek V3 exhibit quite similar performance.
- **Memorization Reliance:** All three models show that the *Verbatim* condition performs nearly identically to the baseline, supporting our hypothesis that models primarily rely on memorization when answering the *Character Guessing* task.
- **Effect of Gist Prompting:** The *Gist* condition results in a non-trivial performance drop, approximately 10%, suggesting that our prompt effectively reduces the model’s dependence on memorization when generating responses.

Our results indicate that gist-based prompting can indeed shift model responses toward reasoning, as evidenced by the performance drop under the *Gist* condition (see Figure 2). However, the moderate performance drop (approximately 10%) suggests that even gist prompting may not completely eliminate the models’ reliance on verbatim memorization. Previous research similarly indicates that prompting without explicit reasoning steps may fail to fully engage true reasoning processes (Yu, 2025). Therefore, to conclusively test our hypothesis—that models’ strong benchmark performance relies significantly on verbatim memorization—we introduce a more direct and explicit intervention aimed at disrupting memorized surface-level cues.

## 5 Restrictive Setting: Testing Reasoning under Memorization Constraints

We further test whether models can reason effectively when direct memorization shortcuts are blocked. In this setting, all character names are replaced with culturally distinct alternatives, removing verbatim recall while preserving narrative meaning. This allows us to evaluate whether LLMs can perform reasoning when surface-level memorization cues are no longer accessible.

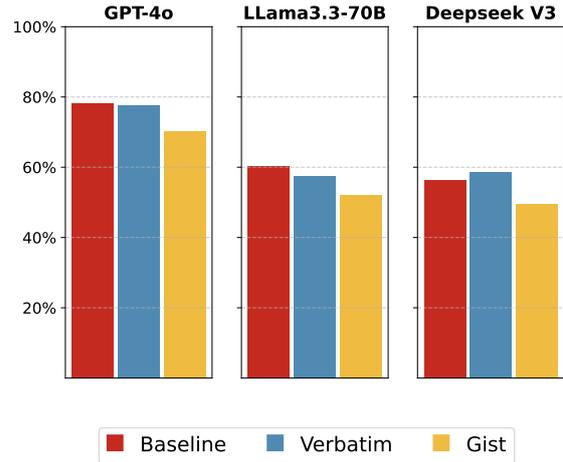


Figure 2: The bar chart compares the performance of three LLMs under three prompting settings. Across all three models, the performance of the Verbatim and Baseline conditions is highly similar, confirming our hypothesis that the models rely on memorization when answering questions. The Gist condition, however, shows a non-trivial decline in performance, indicating that the models struggle when they cannot fully rely on memorization to generate responses.

### 5.1 Name Replacement Strategies and Their Impact on Memorization

**Name Replacement Strategies** Previous studies (Yu et al., 2022) have primarily employed Same-Cultural Name Replacement, a naive and intuitive approach that substitutes character names with culturally consistent alternatives. To systematically analyze the impact of name replacement on model performance, we introduce three distinct levels of name modifications:

- **Name Masking:** Replacing character names with placeholders (e.g., *Sheldon, Amy* → P0, P1).
- **Cross-Cultural Name Replacement:** Substituting names with those from a different cultural background (e.g. English → Chinese) while maintaining gender consistency (e.g., *Sheldon, Amy* → *Jie Zhang, Xiaoling Wang*) (Sun et al., 2024). This tests whether models rely on specific name distributions for character inference.
- **Same-Cultural Name Replacement:** Replacing names with other culturally consistent alternatives (e.g., *Sheldon, Amy* → *Andrew, Sally*). This retains cultural familiarity while disrupting direct memorization by name replacements.

### Human Validation of Reasoning Preservation

We conducted two complementary human validations (Appendix G). *Exp. 1 (Objective accuracy)*

examines whether our name replacement *itself* reduces accuracy; with the accuracy remains no obvious change (max  $|\Delta| = 2.2$  percentage points and no item was judged “unsolvable.”), implying that any model-side drop under replacement is attributable to reliance on verbatim memorization. *Exp. 2 (Gold/semantic integrity)* tests whether replacement changes or ambiguates the gold answer; across 300 item–judgments, none were flagged. Together, these results show that name replacement preserves the semantic structure of the items and keeps gold labels intact.

**Memorization Evaluation Task Format** To directly quantify how much verbatim memorization models rely on, we design a **Source Identification Task**. This task closely resembles established memorization evaluation methods, such as source attribution or textual cloze tests (Chang et al., 2023b). Specifically, given a dialogue excerpt  $D = \{U_1, U_2, \dots, U_n\}$  without source annotations, the model must identify the correct origin of the text, predicting  $f_{\text{source}}(D) \rightarrow S$ , where  $S$  is the specific fictional source.

The rationale behind this design is straightforward: if a model has memorized verbatim text from pretraining, it should accurately identify the source based purely on surface-level textual recall. Conversely, if name replacement or prompting disrupts memorization effectively, we expect a significant decrease in source-identification accuracy. Thus, observing substantial performance degradation would strongly support the claim that models’ prior accuracy was driven largely by verbatim memorization rather than reasoning. We use the **CharScript: Fictional Works Memorization Benchmark Dataset** described in Section 3.2 to rigorously evaluate this effect.

**Impact of Name Replacement on Memorization.** According to Figure 3, our results demonstrate that different name replacement strategies have a significant impact on the memorization ability of LLMs. Among them, Cross-Cultural Name Replacement consistently induces the strongest forgetting effect across all models, indicating that altering names to those from a different cultural background disrupts memorized associations more effectively than other replacement methods. This suggests that cultural distance may play a crucial role in weakening memorized identity-linking patterns in models.

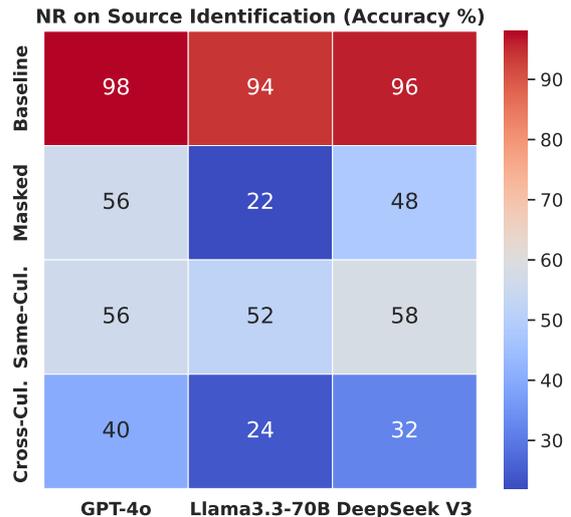


Figure 3: The heatmap presents accuracy percentages for source identification across three LLMs under different name replacement conditions. All three models demonstrate strong memorization of the script. The Cross-Culture Name Replacement method exhibits a significantly strong and the most pronounced memorization-weakening effect across all three models.

## 5.2 Memorization-Controlled Prompting under Name Replacement

Different name replacement strategies have shown a significant impact on the memorization ability of LLMs. Building on this, we apply the name replacement strategies introduced in Section 5.1 to evaluate the *Character Guessing* tasks, using the verbatim and gist prompting strategies we have discussed in Section 4.2.

Under the Name Replacement condition, we provide brief character descriptions generated by GPT-4, aiming to enable the model to complete the Character Guessing task through character matching. The descriptions used are provided in the Appendix C. To assess the actual effectiveness of these descriptions, we conduct an ablation study.

**Results and Key Findings** We present the core results in Figure 4 and summarize four key findings:

- 1. Performance Declines with Reduced Memorization.** Accuracy drops as surface cues are removed, revealing the model’s strong reliance on verbatim memorization.
- 2. Character Descriptions Have Minimal Impact.** Adding character descriptions does not notably change performance, suggesting limited influence on model decisions.

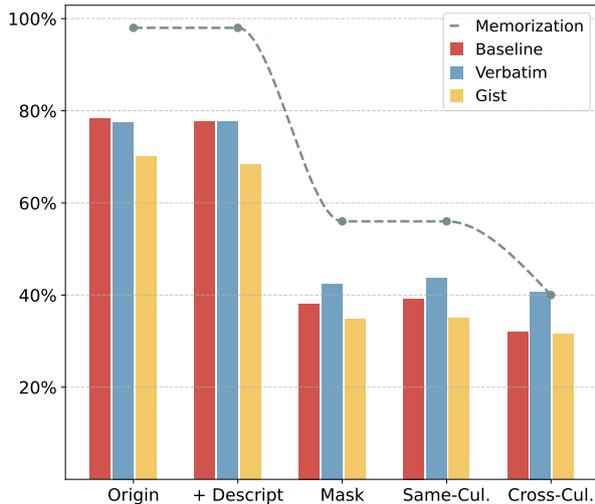


Figure 4: The line chart shows the decline in memorization across different settings, illustrating that as memorization weakens, overall model performance also decreases. The bar chart presents the effect of different prompting strategies: under Name Replacement conditions, Verbatim prompting consistently outperforms the baseline by reinforcing memorization, whereas Gist prompting further reduces reliance on memorization, leading to the lowest accuracy.

3. **Prompting Style Matters.** Verbatim prompts boost performance by reinforcing memorization, while Gist prompts suppress it and lead to lower accuracy.
4. **Name Replacement + Gist Prompting Is Effective.** This combination best reduces memorization reliance and reveals the model’s true reasoning ability.

## 6 Evaluating Understanding under Memorization Constraints

We further evaluate whether our framework generalizes across diverse understanding tasks. Combining inductive prompting with restrictive name replacement, we test if LLMs can sustain reasoning performance when memorization cues are systematically constrained.

### 6.1 Benchmark Tasks Description

We benchmark five types of character understanding tasks spanning coreference, personality inference, role detection, QA, and summarization. Each task evaluates the model’s ability to extract or infer character-related information from dialogue.

**Example Dialogue** To illustrate these tasks, consider the following conversation:

Task	Example Question and Expected Output
<b>Coreference Resolution</b>	"Which entity does 'his' refer to?" → "Grissom"
<b>Personality Understanding</b>	"What trait best describes P1?" → "Analytical"
<b>Role Detection</b>	"Identify all entities referring to the criminal." → "Peter Berglund"
<b>Question Answering</b>	"What play does Grissom mention?" → "Six Degrees of Separation"
<b>Summarization</b>	"Generate a summary." → "Peter discusses convincing a jury, while Grissom references a play."

Table 2: Examples of different character understanding tasks applied to the same dialogue.

**P0:** "You’re still going to have to convince a jury that I killed two strangers for no reason."

*Grissom doesn’t look worried. He takes his gloves off and puts them on the table.*

**P1:** "You ever been to the theater, Peter? There’s a play called Six Degrees of Separation."

The correct speaker labels are **P0** → **Peter Berglund**, **P1** → **Grissom**. For all tasks except Character Guess, the placeholders P0 and P1 are replaced with their original names, Peter Berglund and Grissom. Table 2 illustrates how this dialogue segment is utilized across different tasks.

### 6.2 Gist Prompts Across Character Understanding Tasks

#### General Guidelines for Gist Prompting

- **Encourage Structural Reasoning:** Guide models to infer character identity and relationships from discourse patterns rather than relying on explicit name associations.
- **Eliminate Memorization Bias:** Ensure prompts emphasize evidence-based inference rather than knowledge retrieval.
- **Preserve Narrative Coherence:** In summarization and question answering, focus on maintaining logical flow rather than phrase-matching. Detailed prompts in Appendix F.

### 6.3 Results and Key Findings

Our evaluation provides new insights into how LLMs adapt their reasoning under memorization constraints, summarized in three key findings:

1. **Models Default to Verbatim Recall but Can Adapt under Constraints.** As shown in Table 3, restricting surface cues through name

Tasks	GPT-4o			LLaMA3.3-70B			DeepSeek V3		
	Origin	NR	NR+GIST	Origin	NR	NR+GIST	Origin	NR	NR+GIST
Tvshow Guess	78.2	32.0 ↓46.2	31.5 ↓0.5	61.4	31.9 ↓29.5	27.8 ↓4.1	60.8	29.7 ↓31.1	26.7 ↓3.0
Coreference	58.3	49.6 ↓8.7	47.5 ↓2.1	56.7	48.3 ↓8.4	48.1 ↓0.2	47.4	34.6 ↓12.8	32.3 ↓2.3
FriendsQA	44.3	38.6 ↓5.7	38.6 ↑0.0	45.7	42.9 ↓2.8	44.3 ↑1.4	42.8	39.4 ↓3.4	34.3 ↓5.1
ScreenSum	37.5	28.7 ↓8.8	23.1 ↓5.6	40.1	25.9 ↓14.2	21.3 ↓4.6	35.6	19.3 ↓16.3	17.1 ↓2.2
CSI Role Extract	50.1	32.9 ↓17.2	31.7 ↓1.2	48.3	34.7 ↓13.6	32.7 ↓2.0	45.1	29.8 ↓15.3	23.7 ↓6.1
PERSONET	56.0	52.0 ↓4.0	54.0 ↑2.0	62.0	46.0 ↓16.0	44.0 ↓2.0	44.0	40.0 ↓4.0	38.0 ↓2.0

Table 3: Ablation study examining the impact of name replacement and the GIST method on model performance across six NLP tasks. The subscripted values in the **NR** column indicate the performance drop relative to the original script, while the subscripted values in the **NR+GIST** column show the performance change compared to the Name Replace condition.

replacement consistently reduces accuracy, revealing that models initially rely on verbatim recall. However, when guided by inductive prompting, performance partially recovers, suggesting that LLMs can adjust toward reasoning supported by bounded, context-relevant memorization rather than full recall.

- Task Type Influences Memorization Dependence.** Action-oriented tasks (e.g., Character Guessing, CSI Role Detection) depend more on direct cue-label associations, making them sensitive to memorization constraints. In contrast, motivation-oriented tasks (e.g., Personality Understanding) encourage inference about internal states and intentions, naturally aligning with reasoning that integrates but does not overuse memory.

- Excessive Memorization Can Impair Reasoning.** In some benchmarks, models perform better once memorization is constrained, as removing misleading associations alleviates knowledge conflicts. This indicates that controlled memory use—rather than maximal recall—supports more faithful character understanding.

These findings are consistent across multiple models (Figure 5), suggesting generalizability and broader implications of memorization biases in current benchmarks.

## 7 Conclusion

We examined how LLMs use memorization in memorization-constraint tasks, framing it not as a flaw but as a process that should be constrained and integrated with reasoning. Our two-tier framework—the **inductive** and **restrictive** settings—tests whether models can regulate memo-

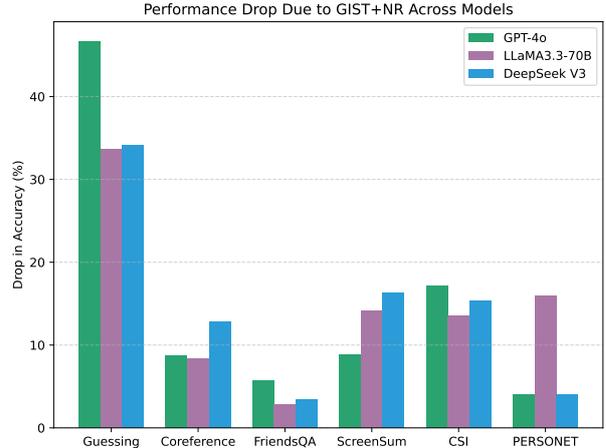


Figure 5: Performance degradation across LLMs due to GIST+NR. Action prediction tasks (e.g., *Guessing*, *CSI*) show greater drops than motivation prediction tasks (e.g., *PERSONET*, *FriendsQA*), reflecting their differential reliance on memorization.

rization when guided by prompts or deprived of surface cues. Results across six character-centric benchmarks show that while models default to verbatim recall, they can be steered toward reasoning under bounded memory. This highlights the need for evaluations that assess reasoning under realistic memorization constraints.

## 8 Acknowledgments

We wish to thank the anonymous reviewers for their helpful comments, feedback, and suggestions. Some experiments were conducted on the UMBC HPCF, supported by the National Science Foundation under Grant No. CNS-1920079. This material is also based on research that is in part supported by DARPA for the SciFy program under agreement number HR00112520301. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copy-

right notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of DARPA or the U.S. Government.

## 9 Limitations

The specific training data of GPT-4o is not publicly disclosed by OpenAI, making it fundamentally unknowable. Our work employs probabilistic inference to estimate the model’s familiarity with a set of fictional works. However, whether these works are explicitly included in the training data remains unanswerable. Nevertheless, even mere familiarity can suffice to contaminate test data, posing challenges to fair evaluation.

Our study can be extended by future research in a number of ways. First, though we ground and motivate our work through “verbatim” and “gist” memorization, we note that any apparent “reasoning” of AI systems may differ fundamentally from those of humans, and we did not pursue human studies to examine this in detail. Further investigation is needed to explore this topic.

Second, the relationship between memorization and reasoning remains a subject of debate in cognitive science. Future studies may uncover insights into this relationship from entirely new perspectives.

Thirdly, While our interventions reveal strong trends, we acknowledge that results may be sensitive to the surface form of prompts. Future work could complement our findings with uncertainty-based analyses such as log probability comparisons.

Finally, our study was centered on select, recent (but not contemporaneous) majority Western popular culture references. Given their popularity, this selection likely increases the chances that LLMs have been pretrained on more instances or language discussing those references (vs. contemporaneous references, or references from non-Western cultures, where overall occurrences in training data may be lower); it is possible that this exaggerated some of the effects we observed.

## 10 Ethics

Our study probes the degree to which GPT-4o have memorized fictional works, and its impact on character understanding tasks. This work uses the OpenAI API on experiments above, and at no point do

we access, or attempt to access, the true training data behind these models, or any underlying components of the systems. **Risks** The several datasets and fictional work scripts in our experiment are sourced from publicly available sources. However, we cannot guarantee that they are free from socially harmful or toxic language. We use ChatGPT<sup>1</sup> to correct grammatical errors in this paper.

**Annotator Characteristics.** The three participants were adult graduate students located in the United States. Two had no prior exposure to the target TV series, while one was familiar with it. All volunteers, no payment. No additional demographic attributes were collected. Instructions given can be found in Appendix G.

**Consent and Data Use.** All participants were informed that the study involved answering comprehension questions for research purposes. They were told that their responses would be anonymized, used only for validation analysis, and not shared beyond the scope of this paper. Written consent was obtained before participation.

## References

- Meta AI. 2024. Llama 3.3 model card. [https://github.com/meta-llama/llama-models/blob/main/models/llama3\\_3/MODEL\\_CARD.md](https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md). Accessed: 2025-02-09.
- Keivan Alizadeh, Seyed Iman Mirzadeh, Dmitry Belenko, S Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. 2024. Llm in a flash: Efficient large language model inference with limited memory. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12562–12584.
- David Bamman, Olivia Lewke, and Anya Mansoor. 2020. *An Annotated Dataset of Coreference in English Literature*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.
- Sabyasachee Baruah and Shrikanth Narayanan. 2023. Character coreference resolution in movie screenplays. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10300–10313.
- Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu

<sup>1</sup><https://chatgpt.com/>

- Purohit, and Edward Raff. 2024. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36.
- Terra Blevins and Luke Zettlemoyer. 2022. [Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models](#). *arXiv preprint*. ArXiv:2204.08110.
- Gordon H Bower and Daniel G Morrow. 1990. Mental models in narrative comprehension. *Science*, 247(4938):44–48.
- 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](#). *arXiv preprint*. ArXiv:2109.05438 [cs].
- Charles J Brainerd and Valerie F Reyna. 2002. Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in child development and behavior*, 28:41–100.
- Nicholas Carlini, Daphne Ippolito, Katherine Lee, Florian Tramèr, Matthew Jagielski, and Chiyuan Zhang. 2023. QUANTIFYING MEMORIZATION ACROSS NEURAL LANGUAGE MODELS. *arXiv preprint arXiv:2202.07646*.
- Tommaso Caselli and Piek Vossen. 2017. [The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction](#). In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.
- Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023a. [Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.
- Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023b. [Speak, memory: An archaeology of books known to chatgpt/gpt-4](#). *arXiv preprint arXiv:2305.00118*.
- Bowen Chen, Namgi Han, and Yusuke Miyao. 2024. A multi-perspective analysis of memorization in large language models. *arXiv preprint arXiv:2405.11577*.
- Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. [SummScreen: A Dataset for Abstractive Screenplay Summarization](#). *arXiv preprint*. ArXiv:2104.07091 [cs].
- Yu-Hsin Chen and Jinho D Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 90–100.
- Lea Frermann, Shay B. Cohen, and Mirella Lapata. 2018. [Whodunnit? Crime Drama as a Case for Natural Language Understanding](#). *Transactions of the Association for Computational Linguistics*, 6:1–15. Place: Cambridge, MA Publisher: MIT Press.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. [News Summarization and Evaluation in the Era of GPT-3](#). *arXiv preprint*. ArXiv:2209.12356 [cs].
- Yuxuan Jiang and Francis Ferraro. 2026. Scribe: Structured mid-level supervision for tool-using language models. *arXiv preprint arXiv:2601.03555*.
- Yuxuan Jiang, Dawei Li, and Frank Ferraro. 2025. [Drp: Distilled reasoning pruning with skill-aware step decomposition for efficient large reasoning models](#). *arXiv preprint arXiv:2505.13975*.
- Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2024. [Disentangling memory and reasoning ability in large language models](#). *arXiv preprint arXiv:2411.13504*.
- Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. [Generalization through Memorization: Nearest Neighbor Language Models](#). *arXiv preprint*. ArXiv:1911.00172.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2024. [Causal estimation of memorisation profiles](#). *arXiv preprint arXiv:2406.04327*.
- Bo Li, Qinghua Zhao, and Lijie Wen. 2024. [Rome: Memorization insights from text, logits and representation](#). *Preprint*, arXiv:2403.00510.
- Dawei Li, Hengyuan Zhang, Yanran Li, and Shiping Yang. 2023. [Multi-level contrastive learning for script-based character understanding](#). *arXiv preprint arXiv:2310.13231*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. [Deepseek-v3 technical report](#). *arXiv preprint arXiv:2412.19437*.
- Robert McKee. 2010. *Story: style, structure, substance, and the principles of screenwriting*. Harper Collins.
- OpenAI. 2023. [Chatgpt: Optimizing language models for dialogue](#). <https://openai.com/research/chatgpt>. Accessed: 2025-02-09.

- Alison H Paris and Scott G Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.
- Akshara Prabhakar, Thomas L Griffiths, and R Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. *arXiv preprint arXiv:2407.01687*.
- USVSN Sai Prashanth, Alvin Deng, Kyle O’Brien, Jyothir SV, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, et al. 2024. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*.
- Valerie F Reyna and Charles J Brainerd. 1998. Fuzzy-trace theory and false memory: New frontiers. *Journal of experimental child psychology*, 71(2):194–209.
- Yisi Sang, Xiangyang Mou, Jing Li, Jeffrey Stanton, and Mo Yu. 2022a. A survey of machine narrative reading comprehension assessments. *arXiv preprint arXiv:2205.00299*.
- Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022b. Tvshowguess: Character comprehension in stories as speaker guessing. *arXiv preprint arXiv:2204.07721*.
- Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022c. **TVShowGuess: Character Comprehension in Stories as Speaker Guessing**. *arXiv preprint*. ArXiv:2204.07721 [cs].
- Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2024. **RoleEval: A Bilingual Role Evaluation Benchmark for Large Language Models**. *arXiv preprint*. ArXiv:2312.16132 [cs].
- Dominik Stammach, Maria Antoniak, and Elliott Ash. 2022. **Heroes, Villains, and Victims, and GPT-3: Automated Extraction of Character Roles Without Training Data**. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. **ConflictBank: A Benchmark for Evaluating the Influence of Knowledge Conflicts in LLM**. *arXiv preprint*. ArXiv:2408.12076.
- Renliang Sun, Mengyuan Liu, Shiping Yang, Rui Wang, Junqing He, and Jiaying Zhang. 2024. Fostering natural conversation in large language models with nico: a natural interactive conversation dataset. *arXiv preprint arXiv:2408.09330*.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelwagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. **MovieQA: Understanding Stories in Movies through Question-Answering**. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, Las Vegas, NV, USA. IEEE.
- Yongqi Tong, Yifan Wang, Dawei Li, Sizhe Wang, Zi Lin, Simeng Han, and Jingbo Shang. 2023. Eliminating reasoning via inferring with planning: A new framework to guide llms’ non-linear thinking. *arXiv preprint arXiv:2310.12342*.
- Sai Vallurupalli, Katrin Erk, and Francis Ferraro. 2024. Saga: A participant-specific examination of story alternatives and goal applicability for a deeper understanding of complex events. *arXiv preprint arXiv:2408.05793*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. **Resolving Knowledge Conflicts in Large Language Models**. *arXiv preprint*. ArXiv:2310.00935.
- Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *arXiv preprint arXiv:2410.23123*.
- Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.
- Ningning Xu, Yuxuan Jiang, Shubhashis Roy Dipta, and Zhang Hengyuan. 2025. Learning how to use tools, not just when: Pattern-aware tool-integrated reasoning. *MATH-AI @ NeurIPS 2025*.
- Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023a. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908.
- Shiping Yang, Jie Wu, Wenbiao Ding, Ning Wu, Shining Liang, Ming Gong, Hengyuan Zhang, and Dongmei Zhang. 2025. Quantifying the robustness of retrieval-augmented language models against spurious features in grounding data. *arXiv preprint arXiv:2503.05587*.
- Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2023b. Harnessing the power of large language models for natural language to first-order logic translation. *arXiv preprint arXiv:2305.15541*.

Zhengzhe Yang and Jinho D. Choi. 2019. [FriendsQA: Open-Domain Question Answering on TV Show Transcripts](#). In *Proceedings of the 20th Annual SIG-Dial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023a. [Personality Understanding of Fictional Characters during Book Reading](#). *arXiv preprint ArXiv:2305.10156* [cs].

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023b. [Personality understanding of fictional characters during book reading](#). *arXiv preprint arXiv:2305.10156*.

Mo Yu, Qiujing Wang, Shunchi Zhang, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Liyan Xu, Jing Li, Yue Yu, et al. 2022. [Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind](#). *arXiv preprint arXiv:2211.04684*.

Yijiong Yu. 2025. [Do llms really think step-by-step in implicit reasoning?](#) *Preprint*, arXiv:2411.15862.

Hengyuan Zhang, Shiping Yang, Xiao Liang, Chenming Shang, Yuxuan Jiang, Chaofan Tao, Jing Xiong, Hayden Kwok-Hay So, Ruobing Xie, Angel X Chang, et al. 2025. [Find your optimal teacher: Personalized data synthesis via router-guided multi-teacher distillation](#). *arXiv preprint arXiv:2510.10925*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. [Don't Make Your LLM an Evaluation Benchmark Cheater](#). *arXiv preprint ArXiv:2311.01964*.

## A Example of Script-based Character Understanding Task

In Table 4 we show a full example of the character understanding task. In this example, we have automatically replaced the individual character names with placeholders like P0 and P1.

## B Extended Dataset Analysis

For the six representative character understanding tasks studied in our work, we selected one dataset for each category. However, we also recognize that there are many different benchmark datasets for each type of task. When selecting representative datasets for each category, we focused on choosing those with more diverse sources. Additionally, we supplemented the selection with eight datasets of the same type but derived from different sources in Table 6.

## C Prompts for GPT4

The following is the prompt we sent to GPT4: **For personality traits extraction** 'you are a psychology research assistant designed to help analysis character personalities according to conversations.' **conversation** 'Here ends the conversation.Give me a character description for each main character according to this conversation.'

**Character Description** Here is a character description:Monica: Responsible, caring, and organized. Assertive and confident in her actions.Chandler: Witty and self-deprecating with an approachable sense of humor. Exhibits insecurity and anxiety in his dialogues, making references to uncomfortable situations and questioning his own actions.Rachel: Spontaneous and open to change, she takes risks and is adaptable. She is also reliant on her relationships with others, signifying her dependency and need for support.' 'Joey: Energetic, extroverted, and casual. Lacks the sensitivity of others feelings at times but genuinely care about friends.Phoebe: Quirky, eccentric, and a free spirit. Her train of thought tends to lean toward the unusual and bizarre. However, she is also compassionate and caring.Ross: Insecure and somewhat neurotic and vulnerable. His behavior is indicative of someone who is going through emotional turmoil.'

## D Name Replacement and Character Comparison

Here is a single example:

Original:

P1: Phoebe, honey, if you hate it so much...

Cross-culture replacement:

P1: Xiaomei, honey, if you hate it so much...

Same-culture replacement:

P1: Susan, honey, if you hate it so much... Neutral placeholder replacement:

P1: Speaker A, honey, if you hate it so much...

We also provide partial name replacement dictionary in table 7.

### D.1 Cross-Language Uncommon Name Replacement

With gender swap:[Bojing, Cuixia, Jingjing, Yunsheng, Meilin, Yusong] Without gender swap:[Cuixia, Bojing, Yunsheng, Jingjing, Yusong, Meilin]

Input	P0: Hey, sorry about that P1: No, we're sorry. We never should have been comparing relationships in the first place. P2: Why? We won. You know, I say, next, we take on Koothrappali and his dog. Really give ourselves a challenge. P3: I just want to say one more thing about this. Just because Penny and I are very different people does not mean that we're a bad couple. P2: The answer is one simple test away. Hmm? You know, it's like when I thought there was a possum in my closet. Did I sit around wondering? No, I sent Leonard in with a pointy stick and a bag. P3: I killed his Chewbacca slippers. P0: Let's just take the test. P3: No, no, no, I don't want to. P0: Oh, well, 'cause you know we're gonna do bad. P3: Because it doesn't matter. I don't care if we're a ten or a two. P2: Or a one. A one is possible. P3: Marriage is scary. You're scared, I'm scared. But it doesn't make me not want to do it. It, it just makes me want to hold your hand and do it with you. P0: Leonard. P1: It makes me so happy if you said things like that. P2: We got an eight-point-two. Trust me, you're happy.
Label	P0: Penny, P1: Amy, P2: Sheldon, P3: Leonard

Table 4: An example from character guessing task.

Input	Tell me the source for following script:Penny: Hey, sorry about that Amy: No, we're sorry. We never should have been comparing relationships in the first place. Sheldon: Why? We won. You know, I say, next, we take on Koothrappali and his dog. Really give ourselves a challenge. Leonard: I just want to say one more thing about this. Just because Penny and I are very different people does not mean that we're a bad couple.
Source	The Big Bang Theory

Table 5: An example for source prediction task.

Task Name	Task Type	Source
Character Mining	Character Identification	TV Shows
LISCU	Character Identification	Book Reviews
MovieQA	Visual Question Answering	Movies
Liter Coreference	Character Coreference	Novels
HVV Detect	Role Detection	Movies, News
SYMON	Visual Question Answering	Movies
CEB	Character Representation	Novels
Love and Violence	Summarization	TV Shows

Table 6: Overview of 8 more tasks in the character understanding benchmark.

Task name	Before replacement	After replacement
TVShow	Monica, Joey, Chandler, Phoebe, Ross, Rachel	Bojing, Cuixia, Jingjing, Yunsheng, Meilin, Yusong
CSI	Grissom, Catherine, Nick, Sara, Brass	Jingbo, Jingwen, Huaqiang, Lihui, Weimin
PERSONET	Villefort, The Count, Dantès, Caderousse, Danglars, Albert, Mercédès, Elders, Abbé Busoni, Eugénie, Faria priest, Ferdinand, Franz, Faria, Morre, Monte Cristo	Weina, Nüxia, Danqing, Cailing, Danyu, Ailan, Mengzhe, Lianhua, Busu, Yujun, Fanli, Feiyan, Fangqing, Fangli, Moya, Mingxia

Table 7: Besides the main characters, there are many names of minor characters and passersby appearing in the tasks. In each task, these secondary and passerby names can be dozens of times more than the number of main characters. Replacing these names is complex but not impossible (e.g., with the help of LLMs). However, to align with the TV show approach of only replacing the names of main characters, we only replaced the main characters in all tasks. From the results, it is clear that even replacing only the main characters has had a significant effect.

replace all the character name [**Monica, Joey, Chandler, Phoebe, Ross, Rachel**] in the conversation with [**Bojing, Cuixia, Jingjing, Yunsheng, Meilin, Yusong**].

#### **The comparison before and after name substitution**

The scene overall reflects a group dynamic characteristic of close friends who are comfortable jesting with each other, while also being there in times of crisis or emotional turmoil.

1. Monica: Responsible, caring, and organized. Assertive and confident in her actions.

1. Bojing: Bojing comes across as practical, level-headed and caring. He often acts as the voice of reason for his friends, attempting to mediate, clarify, and console in various situations. His attempts to play down his date suggest he is a private person who doesn't enjoy sharing intimate details of his life.

2. Joey: Energetic, extroverted, and casual. Lacks the sensitivity of others' feelings at times but genuinely care about friends.

2. Cuixia: Cuixia seems like a lively, fun-loving character. However, she sometimes shows a more cynical side, quick to suspect something might be wrong with Bojing's date and suggesting a strip joint as a solution to Meilin's woes.

3. Chandler: Witty and self-deprecating with an approachable sense of humor. Exhibits insecurity and anxiety in his dialogues, making references to uncomfortable situations and questioning his own actions.

3. Jingjing: Jingjing can be open and candid about his thoughts, even if they seem inappropriate or unusual. He's also humorously self-aware, undercutting his moments of honesty with reminders that he might be oversharing.

4. Phoebe: Quirky, eccentric, and a free spirit. Her train of thought tends to lean toward the unusual and bizarre. However, she is also compassionate and caring.

4. Yunsheng: Appears offbeat and unusual, suggesting the eating chalk anecdote about his past relationship. He also believes in new-age concepts like auras, showing a more spiritual side.

5. Ross: Insecure and somewhat neurotic and

vulnerable. His behavior is indicative of someone who is going through emotional turmoil.

5. Meilin: Exhibits vulnerability and emotional turmoil, especially regarding his recent divorce. He seems to be fluctuating between hurt, anger, and longing for his past relationship.

6. Rachel: Spontaneous and open to change, she takes risks and is adaptable. She is also reliant on her relationships with others, signifying her dependency and need for support.

6. Yusong: Yusong presents as impulsive, high-strung and somewhat comical in moments of panic. Fleeing her wedding because of a sudden realization shows she can make drastic decisions based on her emotions.

## **D.2 Same-Language Replacement**

With Gender-Matched replacement, we replace all the character name [**Monica, Joey, Chandler, Phoebe, Ross, Rachel**] in the conversation with [**Sally, Andrew, Ethan, Olivia, Benjamin, Lauren**].

With Gender Swap as well, we got [**Andrew, Sally, Olivia, Ethan, Lauren, Benjamin**].

## **D.3 Anonymized Speaker Replacement**

replace all the character name [**Monica, Joey, Chandler, Phoebe, Ross, Rachel**] in the conversation with [**Speaker 0, Speaker 1, Speaker 2, Speaker 3, Speaker 4, Speaker 5**]. **LLM summarized character description**

We use the prompt of "you are a psychology research assistant designed to help analysis character personalities according to conversations.....Here ends the conversation. Give me a character description for each main character according to this conversation." " " and we get such result from TVshow Friends scene1.

## **E Benchmark Datasets**

The six selected benchmarks cover a range of character-centric tasks, designed to test different aspects of character understanding (Table 1):

- **TVShow Character Guessing:** Based on the TVSHOWGUESS dataset, which contains scripts from five popular TV series: *Friends*, *The Big Bang Theory*, *The Office*, *Frasier*, and *Gilmore*

*Girls* (Sang et al., 2022b), this is a multi-choice classification task. The dialogue is anonymized using placeholder identifiers (e.g., P1, P2), and the model must predict the correct speaker for each line, testing its ability to identify characters based on context.

- **Screenplay:** Using screenplays from nine popular movies, such as *Avengers: Endgame* (2019), *Dead Poets Society* (1989), *John Wick* (2014), and *The Shawshank Redemption* (1994) (Baruah and Narayanan, 2023), this is a coreference resolution task. The model is tasked with linking character mentions throughout the text, forming coreference clusters and resolving ambiguities in character references.
- **Personality Understanding:** Using the PERSONET dataset, which comprises approximately 32,000 samples from 33 classic literature books (Yu et al., 2023b), this task involves predicting the most likely personality trait depicted in a given text snippet, along with its preceding context. Formulated as a multi-choice problem with distractor traits, this task evaluates the model’s ability to infer character personalities from descriptive text.
- **CSI Role Detection:** Based on 39 episodes of the TV series *CSI* (Fremann et al., 2018), this sequence labeling task requires the model to identify mentions of the perpetrator in dialogues and scene descriptions, focusing on role detection within crime-related narratives.
- **FriendsQA:** Derived from 1,222 scenes from the first four seasons of *Friends* (Yang and Choi, 2019), this dataset supports an open-domain question answering task. The model is given a dialogue excerpt and a manually crafted question and must extract the specific span in the dialogue that answers the question, testing its comprehension of dialogue context.
- **SUMMSCREEN:** Built from transcript-recap pairs of 91 TV series collected from community sources such as The TV MegaSite, Inc. (TMS) and ForeverDreaming (FD) (Chen et al., 2022), this dataset supports a summarization task. The model is required to generate a coherent summary from dialogue utterances and scene descriptions, capturing key plot points while excluding irrelevant details.

## F Gist Prompting Strategies Across Tasks

We design **Gist Prompting** methods to reduce reliance on memorization and promote inference-based character understanding. Below, we detail the task-specific prompt formulations.

### F.1 Task-Specific Gist Prompts

**Character Guessing** *"Analyze the dialogue and infer the most likely speaker based on their speech patterns, personality traits, and relationships with other characters. Do not rely on direct recall of specific names but rather use reasoning from the provided context."*

**Coreference Resolution** *"Identify which mentions in the text refer to the same character by analyzing linguistic patterns, contextual cues, and discourse coherence. Do not rely on pre-learned entity associations but rather resolve coreference based on narrative structure."*

**Personality Understanding** *"Based on the character’s dialogue and actions, infer their personality traits. Focus on how they interact with others, their tone, and behavioral tendencies. Avoid using prior memorized descriptions of characters and derive insights purely from the given passage."*

**Role Detection** *"Determine whether a character plays a key role in the narrative (e.g., antagonist, victim, or protagonist) by analyzing their function within the story. Identify how their actions and relationships define their role rather than relying on predefined labels."*

**Open-Domain QA** *"Answer the question using only the given passage. Identify relevant information by reasoning over the text and analyzing interactions between characters. Do not rely on external knowledge or prior memorized facts."*

**Summarization** *"Summarize the passage by identifying key events, character interactions, and narrative progression. Construct a coherent summary that captures the essence of the scene without directly copying text segments."*

## G Human Study

While our primary focus is to isolate memorization from reasoning in LLMs, we ran a small-scale human study to verify that our name-replacement strategy does not remove essential reasoning clues

and that gold answers remain valid. We implemented two complementary experiments: Exp. 1 evaluates objective accuracy; Exp. 2 evaluates subjective semantic integrity and gold-label validity.

**Participants.** Three adult participants were recruited: two had never watched the target TV series, and one had prior familiarity. All responses were anonymous; no risks were involved.

**Materials.** A set of 100 multiple-choice items (randomly sampled from our benchmark) was used. Each item contains a short excerpt and four options with a pre-defined gold label.

**Exp. 1: Procedure (Objective Accuracy).** Each participant solved the same 100 items twice: first in the *Original* condition (with canonical character names), and then in the *Name-Replaced* condition (with consistent substitutions). After the second round, participants re-examined the items and flagged any case where they believed name replacement removed necessary reasoning information and thus rendered the item unsolvable.

**Exp. 1: Instructions.** “You will be presented with multiple-choice questions based on short excerpts from a TV script. Please answer each question to the best of your ability. After completing both rounds, review the questions again and mark any cases where the replacement of names made the question unsolvable.”

**Exp. 1: Results.** Table 8 reports accuracy before and after name replacement and the number of flagged items. Performance differences were small for all participants (Person 1:  $\Delta = +0.9$  pp; Person 2:  $\Delta = -2.7$  pp; Person 3:  $\Delta = -7.2$  pp; max  $|\Delta| = 7.2$  pp). No item was flagged as unsolvable.

Participant	Accuracy (Original)	Accuracy (Replaced)	Flagged Items
Person 1 (unfamiliar)	46.2%	47.1%	0
Person 2 (unfamiliar)	60.7%	58.0%	0
Person 3 (familiar)	86.6%	84.4%	0

Table 8: Human study (Exp. 1) results before and after name replacement.

**Exp. 2: Procedure (Semantic/Gold Integrity).** To directly test whether name replacement affects gold-answer validity, participants were shown, for each item, the *Name-Replaced* text alongside the

*gold option*. They then judged whether the replacement (i) *changed* the correct answer or (ii) made the answer *ambiguous* given the excerpt.

**Exp. 2: Instructions.** “For each item, you will see the excerpt with replaced names and the official correct option (gold). Please indicate whether the name replacement changes the correct answer or makes it ambiguous. If neither applies, select ‘no issue.’”

**Exp. 2: Results.** Across  $3 \times 100 = 300$  item-judgments, *no* case was flagged as “changed” or “ambiguous.” This indicates that name replacement neither alters gold labels nor compromises their interpretability.

**Discussion.** Exp. 1 shows that human accuracy is stable under name replacement with no unsolvable items, indicating preserved reasoning clues. Exp. 2 shows that gold answers remain valid and unambiguous after replacement. Together, these results support that our intervention suppresses verbatim memorization while preserving the semantic structure required for inference.

## H Prompt Robustness Under Minimal Paraphrases

**Goal.** To test whether our main conclusions are sensitive to prompt wording, we evaluated *minimal paraphrases* of the two prompting conditions (Verbatim vs. Gist). The aim is to keep task framing and structure unchanged while slightly altering phrasing, and to check whether the accuracy pattern in the main paper persists.

**Design.** For each condition we authored two paraphrastic variants (V1/V2 for Verbatim, G1/G2 for Gist), keeping the shared base instruction: “*The following is a dialogue. Please identify who they are.*”

Choose the correct option.” For every model and variant we ran three trials with the same decoding settings in the main experiments (temperature, top-58.0% tokens), and report mean accuracy over the 844 runs.

**Prompt texts (minimal variants).** **Verbatim-V1:** “*Answer by direct memorization retrieval. Do not infer from traits, interactions, or context. Recall exact names from prior dialogue knowledge. Output the final option only; no explanation.*” **Verbatim-V2:** “*Use literal recall of prior dialogue. Do not rely on reasoning about relationships or events.*”

*Output only the letter of the matching name; no steps.*

**Gist-G1:** *“Decide by analyzing relationships, key events, and traits; match these cues to known characters. Do not directly retrieve who spoke from memory. Output the final option only.”*

**Gist-G2:** *“Rely on semantic cues (who relates to whom, what happened, stable traits). Avoid direct memorization of names or lines. Provide only the chosen option.”*

**Results.** Across all models, the qualitative ordering from the main paper is preserved (Baseline  $\geq$  Verbatim  $\geq$  Gist for GPT4o and LLaMA3.370B; for Deepseek V3, Verbatim remains slightly above Baseline, with Gist lowest). Within-condition differences between paraphrases are small (typically within  $\approx 0.3$ – $0.7$  pp), indicating low prompt sensitivity under minimal wording changes.

Model	Baseline	Verbatim		Gist	
	mean	V1 mean	V2 mean	G1 mean	G2 mean
GPT4o	78.23	77.32	77.74	70.48	69.82
LLaMA3.370B	60.19	57.10	57.62	52.22	51.63
Deepseek V3	56.41	58.30	58.86	49.32	49.87

Table 9: Minimal paraphrase robustness. Numbers are mean accuracy (%) over three runs per variant. The qualitative ordering matches the main results; within-condition differences are small.