TVSHOWGUESS: Character Comprehension in Stories as Speaker Guessing

Anonymous ACL submission

Abstract

We propose a new task for assessing machines' skills of understanding fictional characters in narrative stories. The task, TVSHOWGUESS, builds on the scripts of TV series and takes the form of guessing the anonymous main characters based on the backgrounds of the scenes and the dialogues. Our human study supports that this form of task covers comprehension of multiple types of character persona, including understanding characters' personalities, facts and memories of personal experience, which are well aligned with the psychological and literary theories about the theory of mind (ToM) of human beings on understanding fictional characters during reading. We further propose new model architectures to support the contextualized encoding of long scene texts. Experiments show that our proposed approaches significantly outperform baselines, yet still largely lag behind the (nearly perfect) human performance. Our work serves as a first step toward the goal of narrative character comprehension.

1 Introduction

Stories have two essential elements, plots and characters (McKee, 1997). Character comprehension has been widely recognized as key to understanding stories, by psychology, literary and education research (Bower and Morrow, 1990; Kennedy et al., 2013; Currie, 2009; Paris and Paris, 2003; Dymock, 2007). When reading stories, humans can build mental models for characters based on their persona, which helps people to explain a character's emotional status (Gernsbacher et al., 1998), identity, understand her future behaviors (Mead, 1990), and even make counterfactual inference for her own story for that character (Fiske et al., 1979).

The ultimate goal of character comprehension is to equip machines with these human abilities which has direct practical significance. For example, persona can facilitate story generation (Riedl and Young, 2010) and chatbots building (Mairesse and Walker, 2007; Zhang et al., 2018; Urbanek et al., 2019). More importantly, understanding the persona of a particular person can help chatbots to understand the intention behind this person's language (Bender and Koller, 2020), which can lead to better services and ultimately give AI the ability to empathize. For instance, *Amy*'s last sentence in Figure 1 is a joking braggadocio to remind her boyfriend to value her more. Only when *Sheldon* understood the facts of their relationship as a couple and *Amy*'s temporary show-off mentality could he see her true intentions.

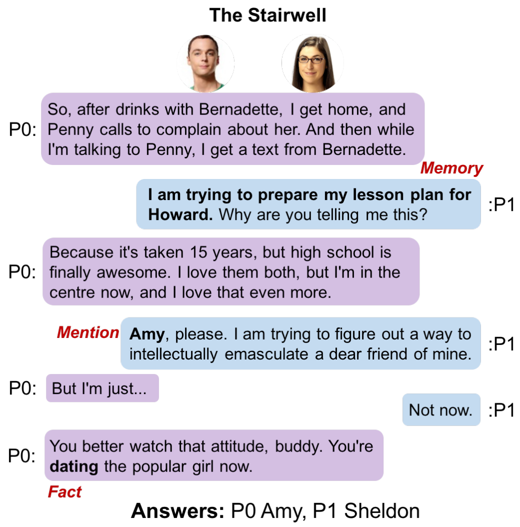


Figure 1: A scene example from TVSHOWGUESS. The character *Amy* can be determined within the scene or with the fact of her relationship; while guessing *Sheldon* would require memory of the character from previous episodes.

Despite the importance, there has been limited attention to modeling characters in stories in the natural language processing (NLP) community.[1] Most existing character-centric prediction tasks have the input sources in expository text such as synopsis (summaries) of stories (Brahman et al., 2021) or non-narrative dialogues (Zhang et al., 2018; Urbanek et al., 2019; Li et al., 2020). A few excep-

[1] In contrast, plot comprehension is a popular NLP topic, especially on event structures (Finlayson, 2012; Elsner, 2012; Sims et al., 2019; Lal et al., 2021; Han et al., 2021).

tions work on stories, but focus on limited aspects of persona, such as facts for coreference resolution (Chen and Choi, 2016), personality (Bamman et al., 2013; Flekova and Gurevych, 2015)and character relationships (Iyyer et al., 2016), with only Chen and Choi (2016); Flekova and Gurevych (2015) provided evaluation benchmarks. Besides the limited persona aspect coverage, they also lack the ability to take into account a theory of mind (**ToM**) which is the knowledge of epistemic mental states that humans use to describe, predict, and explain behavior (Baron-Cohen, 1997).

In this paper, we propose the first task on character comprehension in stories, to assess the ability of mental model construction in NLP. A character's words is her direct reflection to the contexts, conditioned on her character model (Holtgraves, 2010). Our task, **TVShowGuess (TVSG)**, aims to guess anonymous speakers using dialogues, scene descriptions and historical scenes, which requires models to interpret the behavior of characters in the form of dialogues, which meets the requirements for the evaluation of ToMs.

Through experiments and human studies we found: First, the human performance was nearly perfect, while the model performed poorly. Second, although our TVSG has a simple task setup, it has a surprisingly *wide coverage of persona understanding skills* including the linguistic styles, personality types, factoids, personal relations, and the memories of characters' previous experience. Third, most of the cases (>60%) require *identification and understanding of characters' historical experiences* to resolve. Among them, many rely on facts of characters that are not explicitly described in texts but need to be inferred from history events. The wide persona coverage and heavy history dependency challenge existing NLP techniques; and explains the more than 20% accuracy gap between our baselines and humans.

We make the following contributions:

(1) We propose the direction of character comprehension in stories; with an extended survey (Section 2 and Appendix A) discussing the differences and unique challenges compared to related work.

(2) We propose the first task and dataset for this research direction (Section 3).

(3) We propose a new schema to analysis the required evidence for character understanding; and conduct human studies to analyze the required skills of our task (Section 4 and Appendix C).

(4) We propose new model architectures as the initial step of this direction; and conduct comprehensive experiments to provide insights to future work (Section 5 and 6).

2 Related Work

In this section we mainly discuss and compare related work in the two most relevant directions: the assessment benchmarks to the general narrative comprehension skills; and the tasks specifically designed for character-centered predictions over narratives. Table 1 gives a summary of these narrative comprehension tasks, associated with their required skills of comprehension. We also reviewed studies on character-centered tasks over non-narrative texts like synopses and chit-chat (*i.e.*, not story-related) conversations. Detailed rationales of the required skills for each task are discussed in Appendix A.

Assessment of Narrative Comprehension There are many forms of reading comprehension tasks such as cloze tests (Bajgar et al., 2016; Ma et al., 2018), question answering (Richardson et al., 2013; Kočiskỳ et al., 2018; Yang and Choi, 2019; Lal et al., 2021) and text summarization (Ladhak et al., 2020; Kryściński et al., 2021; Chen et al., 2021). Most of these tasks are built on very short stories or can be solved in segments of a story, and therefore present limited challenges to understanding the elements of the story, especially the characters. The exceptions are NarrativeQA (Kočiskỳ et al., 2018) and the three summarization tasks which are mainly event-centric tasks focusing on understanding the plot structures in the stories. The NarrativeQA consists a small portion of character-related questions according to the human study in (Mou et al., 2021), but mainly about simple facts of characters like age, place of birth and profession.

Character-Centric Prediction over Narratives The task of coreference resolution of story characters (Chen and Choi, 2016; Chen et al., 2017a) is most closely related to our TVShowGuess. These tasks focus on identifying the characters mentioned in multiparty conversations, which mainly requires the understanding of discourse relations and assess the personal facts. However, it does not assess the modeling of the character's theory-of-mind, especially the character's memories, as there are no predictions of character behaviors involved. The prediction of fiction characters' personality types by reading the original stories (Flekova and Gurevych, 2015) is another

Dataset	Task Format	Narrative Type		Assessed Narrative Comprehension Skills		
		Source	Length	Plot Structures	Character Facts	Character ToMs
MCTest	Multi-choice QA	Short fiction (Children stories)	~20*	✓		
BookTest	Cloze test	Literature (Excerpt)	–	✓		
(Ma et al., 2018)	Cloze test	TV show transcripts (Scenes)	~20	✓		
NarrativeQA	Generative QA	Movie Scripts, Literature (Full stories)	~11K*	✓	✓	
FriendsQA	Extractive QA	TV show transcripts (Scenes)	~20*	✓	✓	
NovelChapters/BookSum	Summarization	Literature (Chapters or Full stories)	~4K	✓		
SummScreen	Summarization	TV show transcripts (Scenes)	~330	✓		
(Chen and Choi, 2016) / (Chen et al., 2017b)	Coref Resolution	TV show transcripts (Episodes or scenes)	~20/260†	✓	✓	
(Flekova and Gurevych, 2015)	Classification	Literature (Full stories)	~22K		✓	
TVSHOWGUESS	Multi-choice	TV show transcripts (Full stories)	~50K	✓‡	✓	✓

Table 1: Properties of existing narrative comprehension datasets compared to TVSHOWGUESS. * Numbers are not reported in the original paper so we calculated them from the dataset. †(Chen et al., 2017b) proposes two settings: single scene and the whole episode. ‡Our task requires reasoning based on history scenes, which is a form of plot understanding.

character-centric task related to us. These works covers only the personality such as the big five and the MBTI types which is a single perspective of the persona our work considers.

Character-Centric Prediction over Non-Narratives Many tasks do not use the original story, but rather a summary of it. For example, the textual entailment task LiSCU (Brahman et al., 2021) links an anonymous character summary to the name appearing in the story's summary. The usage of summaries prevents the ToM modeling, as discussed in Appendix A.1. Personalized dialogue generation (Mairesse and Walker, 2007; Walker et al., 2012; Zhang et al., 2018; Urbanek et al., 2019; Li et al., 2020) benchmarks are based on daily chit-chats. They usually cover a single aspect of the multi-dimensional persona (Moore et al., 2017), *e.g.*, personal facts (Zhang et al., 2018) or personality types (Mairesse and Walker, 2007; Li et al., 2020). The LIGHT environment (Urbanek et al., 2019) covers both facts and personalities. None of the above covers a comprehensive persona like ours, especially on how a character's past experience builds her ToM.

3 Our TVSHOWGUESS Benchmark

3.1 Task Definition

TVSG adopts a multi-choice setting. The goal is to guess the anonymous speakers who are the main characters (maximum number of 6 for each show) in the scene. The models are provided with an anonymous scene's textual description that consists of n lines $\tilde{\mathcal{S}}^{(t)} = \{\tilde{s}_1^{(t)}, \tilde{s}_2^{(t)}, ..., \tilde{s}_n^{(t)}\}$ (t stands for the t-th scene in the entire show). Each line $\tilde{s}_i$ can be either a dialogue turn or the background description. When the line is a dialogue turn, it is associated with a speaker ID, which can be either the anonymous ID (with the form of P_x, $1 \leq x \leq 6$) of a main character our task studies, or the real name of a supporting character. Similarly, we introduce the notation of the standard scene $\mathcal{S}^{(t)} = \{s_1^{(t)}, s_2^{(t)}, ..., s_n^{(t)}\}$, which has the same definition as the anonymous scenes, with the only difference that the dialogue turns always have their real names of speakers associated.

The anonymous scene $\tilde{\mathcal{S}}^{(t)}$ is associated with a candidate set $\mathcal{C}^{(t)} = c_1^{(t)}, ..., c_k^{(t)}$, $k \leq 6$, with each character $c_j^{(t)}$ is a main character who appears in $\mathcal{S}$. The goal is thus predicting each P_x's actual role $c_j^{(t)}$, *i.e.*, a match $\pi(\cdot)$ from the anonymous IDs to the real characters, conditioned on the scene $\tilde{\mathcal{S}}^{(t)}$ and all the previous scenes $S^{(1:t-1)}$:

$$P(\mathrm{P}_x = c_j^{(t)} | \tilde{\mathcal{S}}^{(t)}, S^{(1:t-1)}) \qquad (1)$$

3.2 Dataset Collection

We collect scenes from the scripts of five popular TV series, including *Friends*, *The Big Bang Theory (TBBT)*, *The Office*, *Frasier* and *Gilmore Girls*.

Data Cleaning Our data consists of character dialogues and backgrounds descriptions. The characters' dialogues start with the characters' names. One or more rounds of dialogue between characters form a scene. Scenes are separated by short backgrounds that begin with markers such as location (e.g. "*Howard's car*", "*Kingman Police Station*"), special words (e.g., "*Scene*", "*Cut*"), or symbols (e.g. "*[ ]*"). To extract information related to our task (i.e., independent scenes) in a structured form, we created a rule-based parser which splits the content of an episode into multiple independent scenes using scene separation markers.

Character Recognition and Anonymization We used main character's names to identify their dialogues within each scene and randomly labeled them as speaker IDs (i.e., P0, P1). Since different names of the characters, such as nicknames, first names and last names, are used in a mixed way to

Show	train	dev	test	#tokens per utterance		#tokens per scene		#tokens per character	
				avg	max	avg	max	avg	max
Friends	2,418	210	211	21	350	862	6,817	190,932	516,191
TBBT	1,791	130	130	19	364	414	6,051	167,027	183,748
Frasier	1,368	140	141	16	363	812	14,276	165,483	475,372
Gilmore_Girls	1,495	141	142	19	336	360	4,572	105,723	214,779
The_Office	3,699	198	199	19	338	123	1,660	58,676	132,992
total	10,771	819	823	18	364	371	14,276	137,568	516,191

Table 2: Statistics of our TVSHOWGUESS.

mark the dialogues. To match lines with the right speakers, we first identified the main characters in each TV show by consulting Fandom's cast lists. Then, we calculated the speaking frequency to find names referring to the same main character.

4 Analysis of Our Benchmark

We propose the first comprehensive **schema of persona types** for the machine narrative comprehension. The schema facilitates the analysis of the challenges in our task; and provides insights of the deficiency in current narrative comprehension models, by allowing a decomposition of model performance to the dimensions of categories (Section 6).

4.1 Our Annotation Schema for Human Study

Two researchers with backgrounds in psychology, linguistics, and education conducted an inductive coding method derived from grounded theory (Glaser and Strauss, 2017). They conducted three rounds of independent annotation and discussion of the evidence needed to identify the characters, using 10 randomly selected scenes for each round. After each discussion, they updated the codebook accordingly. The codebook reached saturation during the process. Then the two researchers coded a total of 318 characters from 105 scenes of Friends and The Big Bang Theory. The annotation interface is attached in Appendix B.

This schema **categorizes the required evidence to resolve the task** into four persona types: *linguistic style*, *personality*, *fact*, *memory*. Table 4 reports inter-rater reliability calculated by Cohen's Kappa (Cohen, 1960). The kappa values are 0.82 for coarse-grained evidence types showing almost perfect agreement (0.81–0.99) (Viera et al., 2005), reflecting the rationality of our scheme.

We also have one additional type *inside-scene*, refers to the tasks that can be resolved within local contexts, thus do not require persona understanding. Furthermore, to better depict how these pieces of evidence are used in human rationales, we added two complementary category scheme: (1) how the task instance **relies on the history scenes** (2) when there are multiple pieces of evidence required, what **types of reasoning skills** are used to derive the answer from the evidence (Section C). Table 6 shows the definitions of each evidence type. We provide examples of each evidence type in Section B.2.

4.1.1 Major Evidence Types

Linguistic style The personalized language patterns which reflect individual differences in self expression and is consistently reliable over time and situations (Pennebaker and King, 1999).

Personality The stable individual characteristics (Vinciarelli and Mohammadi, 2014) which can distinguish "internal properties of the person from overt behaviors" (Matthews et al., 2003).

Memory The character's episodic memory of events from previous episodes and the semantic memory[2] inferred from events.

Fact The truth about characters as opposed to interpretation, which can usually be represented as knowledge triples.

- **Attribute** All explicitly provided factual character identity information in the TV series setting, such as race, occupation, and education level.
- **Relationship** Relationship includes social relationships (e.g., husband and wife) and dramatic relationships (e.g., arch-enemy). When talking to people with different relationships, characters change their identity masks by using different words (Gergen, 1972).
- **Status** The emotional or psychological status of a character when facing a specific situation.

Inside-Scene The textual evidence inside the scene, independent from the characters' persona.

[2]Semantic memory is the characters' general world knowledge that they accumulates over time (Reisberg, 2013). Episodic memory, on the other hand, is the characters' memory of specific experiences in their lives (Tulving, 2002)

- **Background** Background introduction and descriptions in other character dialogues.
- **Mention** The character's name or alias is called by the others. Although mention is persona-independent, it still has challenging cases. Since in a multi-person multi-round chat, common sense of conversational coherence is needed to determine which speaker is being referred to.

Exclusion A guessing technique for elimination using a given list of characters which is neither evidence nor inference, but it depends on the character list provided within the scene, so we include it as a subcategory of inside-scene evidence.

4.1.2 Dependence of History

To understand how much we rely on memory to identify a character, we annotated whether the evidence necessary to solve the task depends directly on historical events or whether it depends indirectly on history by abstracting from historical events.

Direct Dependency Characters that can only be identified through events that are explicitly expressed in previous episodes.[3]

Background: (from `TBBT`) *[The stairwell]*
Candidates: *{Leonard, Penny}*
P0: *There's something I wanted to run past you.*
P1: *What's up?*
P0: *Mm, the guys and I were thinking about investing in Stuart's comic book store. Is that okay?*
P1: *Why are you asking me?*
Answer: P0 → *Leonard*
Rationale: In a previous scene, Leonard and his friends discussed about investing in Stuart's store, so he is the only one between the two who has this memory.

Indirect Dependency Characters can only be identified with evidence that is not explicitly expressed in previous episodes, but can be inferred from previous events. For example, *Personality* can be inferred from the character's previous behavior.[4]

Background: (from `Friends`) *[Central Perk]*
Candidates: *{Joey, Rachel, Ross}*
P0: *Here you are (Hands Rachel a cup of coffee)*
P1: *Thank you Joey. You know what? I'm not even sure I can have caffeine.*
P2: *I went thru this with Ben and Carol. One cup of coffee won't affect your milk.*
P1: *Yeah. Just to be sure I'm gonna call Dr. Wiener.*
Answer: P2 → *Ross*
Rationale: There is not an actual scene on Ross going through this with Carol; the answer is inferred according to Ross' relations to Ben (parent-child) and Carol (ex-spouse). Thus the evidence is facts about Ross and has indirect dependency on the history scenes.

[3]If a character can be identified with evidence of both *Memory* and *Inside-Scene*, it will be labeled as *No-Dependency*.

[4]The annotation of indirect dependency is very subjective as different annotators may have memory of previous scenes and use different evidence to guess the character.

	Evidence Type	Friends(%)	TBBT(%)
(a)	Ling. Style	0.66	9.93
	Personality	7.28	21.85
	Fact	20.53	33.12
	(Attribute)	2.65	8.61
	(Relation)	16.56	22.52
	(Status)	1.32	1.99
	Memory	36.42	27.15
	Inside-Background	33.11	12.58
	Inside-Mention	15.23	15.23
	Exclusion	8.61	22.52

	Dependence of Hist.	Friends(%)	TBBT(%)
(b)	No Dep.	53.64	32.45
	Direct Dep.	26.49	36.42
	Indirect Dep.	19.87	31.13

Table 3: Percentage of the required evidence types in the two TV shows, `Friends` and `The Big Bang Theory`.

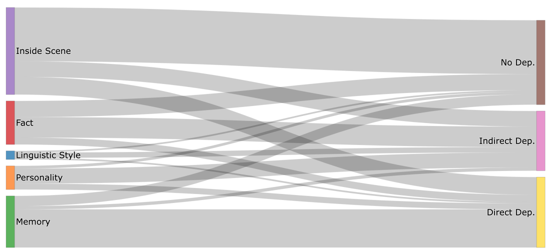


Figure 2: Visualization of the flow from the required evidence types to their dependence of history.

Indirect Dependency If the answer can be inferred within the scene, like answering P0 → *Joey* in the above example. We have a special rule on the *Exclusion* evidence type – If a character can only be inferred on the basis of other characters being solved, it should have dependency type labeled if any of the other character has a history dependency. In other words, when guessing the identity with *Exclusion* requires history dependency on another character, the dependency type is transitive.

4.2 Analysis

Main statistics Table 3 shows the proportions of the required evidence types and dependency of history. According to the statistics, history is an important factor in guessing the characters. 46.36% of the examples from Friends and 67.55% examples from the Big Bang Theory needs history.

Human performance in Accuracy One annotator (who has not watched the evaluating seasons) reports nearly perfect accuracy in guessing the characters in `FRIENDS` (98.68%), and a lower but still good accuracy in `TBBT` (89.82%). A second annotator (who has watched all episodes thus is considered an expert) confirmed that most the error cases are unsolvable given the scenes. We list the unsolvable cases and human mistakes in Appendix E.

Category	$\kappa(\%)$
Evidence type	
Coarse-grained types	81.53
Fine-grained types	80.99
Dependence of history	
Direct dependence only	82.02
All dependency types	75.51
Reasoning Type†	87.21

Table 4: Annotation agreement. †: see our extended study in Appendix C. We list the number for reference.

Correlation between evidence types and history dependence Figure 2 visualizes the flow from evidence types to the dependency of history. Most of them are correlated. Personality and history dependency are most closely related.

5 Methods

Inspired by the successes of applying pre-trained Transformers to reading comprehension tasks, we benchmark our TVSHOWGUESS by building baseline solutions on top of these pre-trained models. The key challenge of our TVSHOWGUESS is that the prediction relies on how a character reacts to the scenario with her/his words, therefore the embedding of each utterance should be highly **context-aware**. This requires to handle the long inputs of scenes, which are usually over the limits of BERT-style models. We propose two solutions. The first is to encode the whole scene with a Transformer with sparse attention (specifically, Longformer (Beltagy et al., 2020)). Then we conduct attentive pooling for each character over the contextualized embeddings of all her utterances. The second is to organize each utterance with its necessary history context (as one row), and have a BERT model to encode each relatively short utterance independently and use an attention module to summarize the rows of the same masked character for final prediction.

5.1 Transformers with Character-Pooling

Our first approach (the top in Figure 3) is denoted as Longformer-Pooling (or **Longformer-P**).

Scene Encoding The input $\tilde{S}$ to the model includes the concatenation of all the utterances in an anonymous scene. Each utterance is prefixed by a speaker ID token and suffixed by a separation token, *i.e.*,

$$T_i = [\mathrm{P}_{x_i}] \oplus U_i \oplus [\mathrm{SPLIT}]$$
$$\tilde{S} = T_0 \oplus T_1 \oplus ... \oplus T_N,$$

where U_i is the i-th utterance and $[\mathrm{P}_{x_i}]$ is its speaker ID (e.g., $[\mathrm{P}_0]$ and $[\mathrm{P}_1]$). [SPLIT] is a special token. $\oplus$ denotes concatenation. We use a Longformer to encode the whole $\tilde{S}$, to make the embedding of each utterance token *context-aware*, *i.e.*, $\mathbf{H} = \mathrm{Longformer}(\tilde{S}) \in \mathbb{R}^{L\times D}$.

Character-Specific Attentive Pooling For each character ID P_x, we have a mask $M_x \in \mathbb{R}^{L\times 1}$ that has value $M_x[j] = 1$ if the j-th word belongs to an utterance of P_x; and 0 otherwise. For each character P_x, we then collect the useful information from all her utterances as masked by M_x as

$$A = \mathrm{Attention}(\mathbf{H}),\ \alpha_x = \mathrm{Softmax}(A \odot M_x).$$

The character-specific attention α_x is then used to pool the hidden states to summarize a character representation in the input scene $\tilde{S}$ and make the prediction: $P(\mathrm{P}_x = c|\tilde{S}) = f_k(\mathbf{H}^T\alpha_x)$. Here $f_k : \mathbb{R}^{d\times 1} \to \mathbb{R}^{C\times 1}$ is the character classifier for the k-th TV show.

5.2 Multi-Row BERT

The second approach (the bottom in Figure 3) is denoted as the multi-row BERT (**MR. BERT**). We split the long scene $\tilde{S}$ into multiple segments $\{\tilde{s}_i\}$. Encoding the segments reduces the overall complexity from $O(L^2)$ to $O(RL_s^2)$, where L_s is the maximum segment length and $L_s \ll L$. For the construction of each segment, we take an utterance T_i in Eq. (2), concatenated with the history utterances $T_{i'} (i' < i)$ until arriving the maximum length L_s. We sample R such segments to make sure each P_x have at least one segment. During sampling we also use a trick to focus more on the end of the scene, as these utterances have more histories so they will cover more contents from the scene (*the reverse trick*).

$$\{\tilde{s}_i\} = \begin{bmatrix} T_{t_1} \oplus [\mathrm{SEP}] \oplus T_{t_1 - 1} \oplus T_{t_1 - 2} \cdots \\ T_{t_2} \oplus [\mathrm{SEP}] \oplus T_{t_2 - 1} \oplus T_{t_2 - 2} \cdots \\ \cdots \\ T_{t_R} \oplus [\mathrm{SEP}] \oplus T_{t_R - 1} \oplus T_{t_R - 2} \cdots \end{bmatrix}.$$

Then we encode the $\{\tilde{s}_i\}$ with a BERT encoder:

$$\mathbf{H} = \mathrm{BERT}(\{\tilde{s}_i\}) \in \mathbb{R}^{R\times L\times D}.$$

Finally, similarly to Longformer-P, we have a mask of rows $M_x \in \mathbb{R}^R$ for each character ID P_x, with $M_x[j] = 1$ if the j-th row is an utterance of P_x. Then we apply the same attentive pooling technique and make the prediction as in Longformer-P.

6 Experiments

6.1 Baselines and Implementation Details

We also compare with the vanilla pre-trained Transformer baseline, **Vanilla Longformer Classifier**.

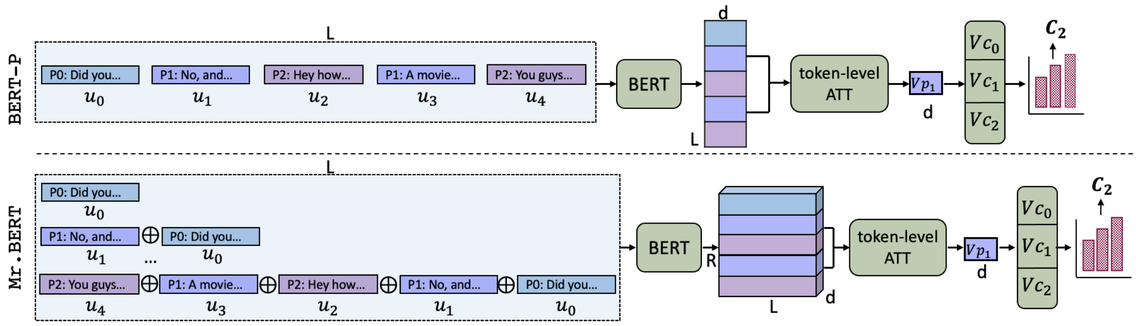


Figure 3: Our two proposed model architectures for the character prediction task.

System	FRIENDS dev	FRIENDS test	TBBT dev	TBBT test	Frasier dev	Frasier test	Gilmore_Girls dev	Gilmore_Girls test	The_Office dev	The_Office test	Overall dev	Overall test
Random	35.23	31.59	33.08	37.79	34.74	31.61	36.43	38.90	44.30	46.71	36.79	36.59
Vanilla Longformer	67.79	60.63	61.58	63.95	85.11	82.06	79.84	74.52	70.92	71.60	72.55	69.72
repl with BERT	65.60	59.58	61.58	58.43	85.11	84.30	81.91	70.41	67.56	68.54	71.65	67.76
Our MR. BERT	**77.01**	**73.20**	62.60	62.50	90.07	82.51	**83.98**	**78.63**	70.92	74.41	76.82	74.52
- context	62.92	57.19	59.54	63.95	81.64	76.23	74.42	67.12	66.00	67.37	68.33	65.54
- reverse trick	70.81	68.71	52.42	59.01	79.40	81.39	78.04	73.97	66.22	68.31	69.45	70.52
- fill-empty trick	74.33	68.56	58.27	63.37	86.10	78.48	72.87	69.86	68.90	73.71	72.28	70.92
Our Longformer-P	**77.01**	69.91	**63.87**	**66.57**	**90.32**	**87.67**	82.17	75.07	**71.81**	**76.29**	**76.95**	**74.97**
maxlen=1000	74.16	66.77	63.36	64.24	86.10	85.65	79.33	72.05	73.83	76.06	75.25	72.74
repl with BERT	68.12	58.83	61.32	63.95	82.63	76.91	68.48	65.75	72.48	71.83	70.49	66.79
Human*	98.68	–	89.82	–	–	–	–	–	–	–		

Table 5: Overall performance (%) on our TVSHOWGUESS task. (*) Human evaluation was conducted on a subset of the dataset.

The model conducts direct classification over the concatenation of a character's utterances in the scene. It can be viewed as a discriminative language model of the characters' lines.

We include the implementation details of the baseline and our models in Appendix G.

6.2 Results

Overall results Table 5 compares different models on our TVSHOWGUESS. Our proposed architectures beat our vanilla character classifier with large margins (4-5%). However, human performance is significantly (21-26%) better than the best models , showing models are still far from reaching human level of character understanding.

Among all the shows, `TBBT` is the most challenging one, while `Frasier` and `Gilmore Girls` are relatively simpler. Given that there is no correlation between performance and scene lengths (Table 2), this shows the difficulties of the tasks mainly come from the persona modeling, inference and reasoning. Specifically, the *Inside-Scene* evidence requires less persona understanding. Therefore, the relatively smaller amount of *Inside-Scene* cases makes `TBBT` more difficult. Also the existing models are not good at resolving the related memory or facts from the history, thus the high ratio of *history dependent* cases in `TBBT` also leads to lower performance.

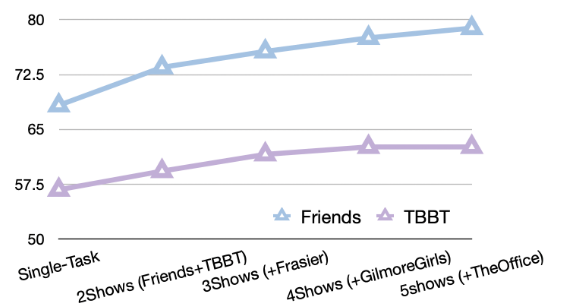


Figure 4: Learning curves of the two TV shows with increasing training data from other shows.

6.3 Analysis

Learning Curves We plots the learning curves of `Friends` and `TBBT`, with increasing number of shows used as training data (Figure 4). The curves become flat with all shows added, showing that our task has sufficiently data for training.

Impact of the dependence on history The bar charts in Figure 5 show the performance on different history dependence types. The performance of cases that require history supports is in general harder for most of our models (~20% lower com-

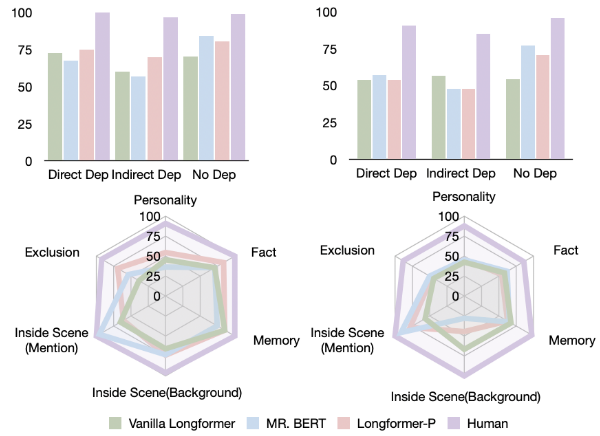


Figure 5: Performance breakdown according to our schema (left: Friends, right: The Big Bang Theory).

pared to the cases without dependency of history).

The results indicate that to further improve the model performance, the models are required to better model the history events associated with each character. This perfects aligns with the theories that past experience is an important fact to build characters' ToM, showing that our TVSHOWGUESS does serve as a good benchmark for the in-depth study of character comprehension from stories.

Another interesting finding is that the cases requiring indirect history dependence (usually *Personality* and *Facts*) are even more challenging. Humans can build a structured profile of characters when reading stories. The neural models represent each character as a single vector (*i.e.*, the weight vector in the output layer), with different items in one's profile mixed. This indicates a promising future direction of constructing structured persona representations (*e.g.*, based on our schema of evidence) for more accurate character modeling.

Breakdown to evidence types The wind-rose charts (bottom) in Figure 5 provide performance breakdown onto our evidence categories. We omit the type of *Linguistic style* because there are only two cases in `Friends` so the results are not stable.

As expected, the cases that can be resolved locally without character understanding (*Inside-Mention*) are relatively easier. All of *Personality*, *Fact* and *Memory* cases have much lower performance as they correspond to heavy dependency on the modeling of history.

The type *Exclusion* gives the worst overall performance on the two shows. However, this does not indicate difficulty of character understanding – According to the definition, these cases cannot be directly resolved with the scene inputs, but require the model to have specific strategy to exclude some incorrect answers first.

It is surprising that the *Inside-Background* type poses difficulties to our models, because it looks to human annotators mostly standard textual inference.[5] We identify two possible reasons: (1) As discussed in the introduction, some cases require pragmatic understanding from the surface form to intention, only on which textual inference can be performed (2) The portion of this type is relatively smaller so the model may fail to recognize the required textual inference skills during training.

Effect of Scene Contexts Finally, the vanilla character classifier has a quite different behavior compared to the other models. Because it cannot make use of contexts within scenes, there is a great drop on the *Inside-Mention* type (hence the drop on the *No Dep* type). However, it does not suffer from significant drop on the other types. This indicates none of the current models have clear advantage on modeling persona; and our task is in general challenging to existing NLP techniques.

Challenges of History Retrieval Our experiments show that the history dependency challenges existing models. Finding the evidence from history scenes is a retrieval task (but without groundtruth). To see how it brings new challenges to existing semantic search, we applied a state-of-the-art model to retrieve the history scenes and conducted an additional human study to evaluate the results. Our study shows that on our identified cases with *Direct Dependency*, the top-3 results (from in total 20 candidates) of a state-of-the-art semantic search model only give a recall of 35.5%. The result confirms that our task requires further advances on semantic retrieval. The detailed setting and our discussions can be found in Appendix F.

7 Conclusion

In this paper, we present the first task and dataset for evaluating machine reading comprehension models for understanding characters in narratives. Based on linguistic, education, and psychology theories, we propose a new schema and conduct two human studies to analyze the types of evidence and reasoning required in understanding characters. We further design a new model architecture and conduct comprehensive experiments to serving as a testbed for future studies.[6]

[5] In NLP community, people usually agree that textual inference is within the realm of pre-trained LMs.

[6] We will release our data and data (under MIT license).

References

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956*.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

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*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL 2017*, pages 1870–1879.

Henry Y Chen, Ethan Zhou, and Jinho D Choi. 2017b. Robust coreference resolution and entity linking on dialogues: Character identification on tv show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Susan Dymock. 2007. Comprehension strategy instruction: Teaching narrative text structure awareness. *The Reading Teacher*, 61(2):161–167.

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644.

Mark Mark Alan Finlayson. 2012. *Learning narrative structure from annotated folktales*. Ph.D. thesis, Massachusetts Institute of Technology.

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

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816.

Kenneth J Gergen. 1972. Multiple identity: The healthy, happy human being wears many masks. *Psychology today*, 5(12):31–35.

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

Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Rujun Han, I Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, Nanyun Pen, et al. 2021. Ester: A machine reading comprehension dataset for event semantic relation reasoning. *arXiv preprint arXiv:2104.08350*.

Thomas Holtgraves. 2010. Social psychology and language: Words, utterances, and conversations.

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

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

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.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. Tellmewhy: A dataset for answering why-questions in narratives. *arXiv preprint arXiv:2106.06132*.

Aaron W Li, Veronica Jiang, Steven Y Feng, Julia Sprague, Wei Zhou, and Jesse Hoey. 2020. Aloha: Artificial learning of human attributes for dialogue agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8155–8163.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Kaixin Ma, Tomasz Jurczyk, and Jinho D Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. 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 2039–2048.

François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 496–503.

Gerald Matthews, Ian J Deary, and Martha C Whiteman. 2003. *Personality traits*. Cambridge University Press.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

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

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

Christopher Moore, Kim Barbour, and Katja Lee. 2017. Five dimensions of online persona.

Daniel G Morrow. 1985. Prominent characters and events organize narrative understanding. *Journal of memory and language*, 24(3):304–319.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative question answering with cutting-edge open-domain qa techniques: A comprehensive study. *arXiv preprint arXiv:2106.03826*.

Isabel Briggs Myers and Mary H McCaulley. 1988. *Myers-Briggs type indicator: MBTI.* Consulting Psychologists Press Palo Alto.

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

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Ijcai*, pages 4279–4285.

Daniel Reisberg. 2013. *The Oxford handbook of cognitive psychology*. Oxford University Press.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

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

Endel Tulving. 2002. Episodic memory: From mind to brain. *Annual review of psychology*, 53(1):1–25.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. *arXiv preprint arXiv:1903.03094*.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Marilyn A Walker, Grace I Lin, Jennifer Sawyer, et al. 2012. An annotated corpus of film dialogue for learning and characterizing character style. In *LREC*, pages 1373–1378.

Zhengzhe Yang and Jinho D Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

A A Detailed Survey of Related Work

We first gave an in-depth analysis on the difference between narrative and synopsis, from both the empirical challenges in NLP studies and the linguistic theory from (Morrow, 1985). Then we provide detailed discussion on how we summarize related work in Table 6.

A.1 Background: Narrative versus Synopsis

As our work focuses on narrative comprehension, following the setups like (Kočiskỳ et al., 2018; Kryściński et al., 2021; Chen et al., 2021), it is necessary to make the difference clear between comprehension of the original narrative stories versus comprehension of their synopses (the human-written plot summaries), *e.g.*, from the story's Wikipedia page.

Narrative stories are told by creating scenes, with the goal of making readers directly experience events as they occur, and empathize with the story characters in relation to their own experiences. To engage the readers, story writers usually use complex narrative clues (*e.g.*, character activities, event development, scenery changes); variable narrative sequence (*e.g.*, narrative, flashback, interpolation); and a variety of expressions (*e.g.*, argument, lyricism, narrative, description, illustration). By comparison, a synopsis is a descriptive summary of the main idea of a story while keeping the language simple. It contains only the main characters, time, place, important plot, and ending, rather than allowing the story to unfold through the actions of the characters. The goal is to inform the readers what happened without much involvement of the original story.

Therefore, comprehension of narrative stories requires more sophisticated skills to understand the complex clues and expressions, in order to finally build a narrative representation from a sequence of scene comprehension and empathize with the characters based on the understanding of their mental models (Morrow, 1985). A synopsis can be regarded as the processed results from the above skills from a (experienced) human reader, thus reducing the major parts of narrative understanding.

A.2 Assessment of Narrative Comprehension

We summarize the related tasks people use for assessment of general narrative comprehension skills.

Cloze Test Cloze tests take a snippet of the original text with some pieces (usually entities) masked as blanks, with the goal of filling these blanks from a list of candidates. The cloze tests can be automatically constructed, resulting in an advantage of easy to get large scale datasets. Examples of cloze tests for narrative comprehension assessments are BookTest (Bajgar et al., 2016) and (Ma et al., 2018). Both datasets are based on excerpts of books or scenes of TV shows. As the machines are only provided with short paragraphs, there are not sufficient information to infer complex character set via reading the stories. Therefore, these datasets cover few questions assessing the understanding of characters.[7]

Moreover, when built on short snippets, the cloze tests is known to prone to mostly local inference but not much reasoning and commonsense knowledge, as pointed by studies in the NLP community suggested (Chen et al., 2016). On the other hand, although our task also has form similar to cloze style, it requires information about the characters from previous stories, which is not only about understanding the characters, but also requires global inference of the story (see Figure 1).

Question Answering The most popular form of narrative comprehension evaluation is through question answering, starting from the early work of MCTest (Richardson et al., 2013), to the more recent crowd-sourced tasks like NarrativeQA (Kočiskỳ et al., 2018), FriendsQA (Yang and Choi, 2019), and TellMeWhy (Lal et al., 2021).

Among them, the MCTest and TellMeWhy conduct multi-choice question answering on short stories. As the machines are only provided with short paragraphs, there are not sufficient information to infer complex character set via reading the stories. Therefore, these datasets cover few questions assessing the understanding of characters. The TellMeWhy has a specific focus on *why*-questions assessing the causal knowledge between states and events. The inputs are short stories from the ROCStories dataset (Mostafazadeh et al., 2016). MCTest covers much wider classes of reading skills, as it bases on complete stories, and generates questions with the goal of assessing

[7] There may be a possible confusion of these tasks and ours, as they also require to fill the anonymous character names in the blanks. However, in these tasks, the required answers are also anonymized character IDs that appear in the inputs, and the IDs for the same character are random across different scenes. Therefore the character's information is not available for learning by design. In other words, their design of tasks *deliberately prevent* the task of character understanding.

Dataset	Task Format	Narrative Type Source	Length	Assessed Narrative Comprehension Skills Plot Structures	Character Facts	Character ToMs	Assessed Commonsense Knowledge Concepts	Events/States	Story Flows
MCTest	Multi-choice QA	Short fiction (Children stories)	∼20*	✓			✓	✓	✓
BookTest	Cloze test	Literature (Excerpt)	-	✓					
(Ma et al., 2018)	Cloze test	TV show transcripts (Scenes)	∼20	✓					
NarrativeQA	Generative QA	Movie Scripts, Literature (Full stories)	∼11K*	✓	✓			✓	
FriendsQA	Extractive QA	TV show transcripts (Scenes)	∼20*	✓	✓				
TellMeWhy	Multi-choice QA	Short fiction (ROCStories)	5					✓	
NovelChapters/BookSum	Summarization	Literature (Chapters or Full stories)	∼4K	✓					✓
SummScreen	Summarization	TV show transcripts (Scenes)	∼330	✓					✓
(Chen and Choi, 2016) / (Chen et al., 2017b)	Coref Resolution	TV show transcripts (Episodes or scenes)	∼20/260†	✓	✓			✓	✓
(Flekova and Gurevych, 2015)	Classification	Literature (Full stories)	∼22K		✓				
TVSHOWGUESS	Multi-choice	TV show transcripts (Full stories)	∼50K	✓(indirect)	✓	✓	✓	✓	✓

Table 6: Properties of existing narrative comprehension datasets compared to TVSHOWGUESS . We organize the datasets according to the following dimensions related to narrative understanding: **Source** of the texts for reading comprehension; **Length** of the texts from the source that makes the task solvable, we report the numbers of sentences or utterances for books and scripts respectively; whether the task assesses the ability of understanding **plot structures** in the stories; whether the task assesses the ability of understanding basic **character facts** like personality, profession, etc; whether the task assesses the ability of building **character theory-of-mind (ToM)**; whether the task assesses the commonsense knowledge of **concepts**, **events** and **states**; and whether the task assesses the additional commonsense about the **narrative development**, including the knowledge about the coherence among non-verbal narratives and dialogues, and how they form the story/plot flow. * Numbers are not reported in the original paper so we calculated them from the dataset. †(Chen et al., 2017b) proposes two settings with single scene and the whole episode as inputs respectively. Different from ours, their include of episode is not to support the in-scene prediction with necessary history, but mostly increase the difficulty level of the co-ref task.

children's reading comprehension over both story plots and commonsense.

NarrativeQA and FriendsQA conduct natural question answering tasks. NarrativeQA aims to infer free-form answers to questions about a specific book or movie script. According to the human study from (Mou et al., 2021), the major part of the dataset is event-centric questions, which queries the explicit plots from the original books thus do not require a significant amount of commonsense reasoning. The study also reveals that NarrativeQA consists of a small portion of character-related questions. These questions mainly query the simple facts of characters, such as age and profession. The more complexity character persona types, like personality, emotional/psychological status and history experience studied in our work, are not covered. Similar to NarrativeQA, FriendsQA is a QA task over TV show scripts. The dataset consists of six types of questions: *who, what, when, where, why*, and *how*. The *who* questions target on asking speaker names of utterance contents or participants of events, therefore are mainly assessing understanding of plot structures (*i.e.*, participant arguments of events).

Both NarrativeQA and FriendsQA have human-written questions with a reference of the plot summary, which require evidence explicitly exists in the original story texts, thus do not have much requirement of reasoning. The FriendsQA questions are based on scene summaries, thus require mostly local evidence; the NarrativeQA questions are based on the book-level summary, thus sometimes require the ability to bridge the gap between coarse-grained and fine-grained event descriptions (*i.e.*, commonsense of sub-events).

Summarization There is a recent trend to evaluate model's understanding of stories via summarization, including NovelChapters (Ladhak et al., 2020), BookSum (Kryściński et al., 2021) and ScreenSum (Chen et al., 2021). These works provide a good research opportunity to future story reading research, by showing that book-level or chapter-level summarization is challenging to existing machine reading models. However, it is more difficulty to identify the specific required reading skills by these tasks, as there exist many factors beyond reading skills to generate a good summary, such as encoding and generating long narrative texts. Intuitively, story summarization is largely plot-related instead of character-related; and requires the knowledge to understand the story flow.

A.3 Character-Centric Prediction over Narratives

Our task can be seen as a character-centered understanding of the narrative, where the understanding of the character deepens the understanding of the story and makes the narrative engaging. There are limited studies on understanding characters' persona from reading stories. In this section we review some existing character-centric prediction tasks over narrative texts, and discuss the relations and differences.

Character Name Linking The task of coreference resolution for story characters (Chen and Choi, 2016; Chen et al., 2017b) is closely related to our TVSHOWGUESS. These coreference resolution focuses on identifying the characters mentioned in multiparty conversations from TV shows scripts. The goal of these tasks is to resolve the coreference of pronouns and character-indicating nominals (*e.g.*, *you* and *Mom*) **in dialogues** of the character names that appear in the local context. It also covers linking a named entity (*e.g.*, *Ross*) to the character, which is more on name matching instead of character understanding.

The task form of coreference resolution mainly requires the understanding of discourse relations. It does not assess the modeling of character theory-of-mind, especially the character's memories, as there are no predictions of character behaviors involved. The major character persona type it assesses is character facts, since the resolution of nominals requires the understanding of the target characters' occupations and relationships.

The lack of ToM modeling and complex reasoning of the coreference resolution task also makes it relatively easier – on `Friends` and `The Big Bang Theory`, a CNN model gives a >90% average accuracy. By comparison, our task, although solvable by humans with a ~95% accuracy, is challenging to neural models as the best BERT-based model gives a ~65% average accuracy on the same two shows with even smaller candidate sets.

Personality Prediction Our work is also related to the prediction of fiction characters' personality types by reading the stories (Flekova and Gurevych, 2015). Specifically, the tasks require to predict a fiction character's MBTI personality types (Myers and McCaulley, 1988) rooted from Jung's theory, based on the character's verbal and non-verbal narratives in the original stories. Compared to the aforementioned character-centric prediction tasks, these studies require to read and comprehend the original long stories, but the prediction task are relatively simpler since they only focus on personality which is a single perspective of persona.

A.4 Character-Centric Prediction over Non-Narratives

Character name linking between story synopses Recently Brahman et al. (2021) propose the LiSCU, which is a novel textual entailment task linking an anonymous summative descriptions of story character to the name appearing in the story's plot summary. Similarly to (Chen and Choi, 2016), the task assess the resolution of names and events instead of the ToM modeling. This is because the task does not involve much explicit behavior predictions, since the task form is entailment between two given statements rather than predicting the possibility of new contents. The usage of synopses over original stories reduces the challenges in narrative understanding; and further prevents the character comprehension from stories, as pointed out by (Kočiskỳ et al., 2018), the summaries themselves are humans' comprehension results of the stories.

Personalized Dialogue Generation Finally, our work is also related to personalized dialogue generation, for which datasets (Mairesse and Walker, 2007; Walker et al., 2012; Zhang et al., 2018; Li et al., 2020) and models (Li et al., 2016; Mazaré et al., 2018; Qian et al., 2018; Zheng et al., 2020) are proposed for generating dialogues for speakers with persona features. These benchmarks usually cover a single aspect of the multi-dimensional persona (Moore et al., 2017). For example, `PERSONA-CHAT` (Zhang et al., 2018) focuses on personal facts such as "*I'm a writer*" and "*I live in Springfield*"; other works mainly focus on learning language styles from speakers' personality types, such as the Big Five traits of the extraversion personality in `PERSONAGE` (Mairesse and Walker, 2007), and the personality types derived from TV tropes (e.g. *jealous girlfriend*, *book doom*, *anti hero*) in `ALOHA` (Li et al., 2020).

`LIGHT` (Urbanek et al., 2019) is a crowd-sourced dataset for text game adventure research. It includes natural language descriptions of fantasy locations, objects and their affordances, characters and their personalities, dialogue and actions of the characters. The biggest difference between ours and LIGHT is that LIGHT is based on the local environment of the conversation, rather than on a story. Examples from the LIGHT dataset are independent conversations and the context in which they occur. Crowd workers created the dialogues of characters by a given setting and a persona. The persona is modeled by the Persona-Chat dataset which is defined as a set of three to five profile sentences describing their personal facts such as "*I am a part of a group of travelers*" and "*I go from town to town selling food to the locals*".

To the best of our knowledge, none of the existing studies cover a comprehensive multi-dimensional persona like in our work, especially on how a character's past experience builds her ToM.

B Supplementary for the Dataset Analysis

B.1 Summary of the Annotation Schema

We include a summary of our annotation schema in Figure 6.

B.2 Examples of each evidence types

Linguistic style

Background: (from TBBT) *[Amy's car]*
Candidates: *{Leonard, Penny, Sheldon, Amy}*
P0: *Whatever. You can't even go on a date without checking your relationship agreement.*
P1: *If you've got a problem basing a relationship on a contract, I'd like to tell you about 13 plucky colonies that entered a relationship agreement called the U.S. Constitution. And it may not be cool to say so, but I think that love affair is still pretty hot today.*
Answer: P1 → *Leonard*
Rationale: (Shelton's language is characterized by the use of long, difficult sentences and references to historical stories.)

Personality

Background: (from TBBT) *[The cafeteria]*
Candidates: *{Leonard, Howard, Sheldon, Raj}*
P0: *And you love the sound of your own voice.*
P1: *Yeah, well, of course I do. Listen to it. It's like an earful of melted caramel.*
Answer: P1 → *Sheldon*
Rationale: (Sheldon is a self-centered person so he will praise his own voice.)

Memory

Background: (from TBBT) *[The stairwell]*
Candidates: *{Leonard, Penny}*
P0: *There's something I wanted to run past you.*
P1: *What's up?*
P0: *Mm, the guys and I were thinking about investing in Stuart's comic book store. Is that okay?*
P1: *Why are you asking me?*
Answer: P0 → *Leonard*
Rationale: (In a previous scene, Leonard and his friends discussed about investing in Stuart's store, so he is the only one between the two who has this memory.)

Fact

- Attribute

Background: (from TBBT) *[Amy's lab]*
Candidates: *{Amy, Penny}*
P0: *Hey. Ready to go to lunch?*
P1: *Just give me a minute. I'm stimulating the pleasure cells of this starfish. I just need to turn it off.*
Answer: P1 → *Sheldon*
Rationale: (Sheldon is Amy's boyfriend. After identify P0 is Amy, based on the relationship between Amy and Sheldon, P1 can be identified as Sheldon.)

- Relationship

Background: (from TBBT) *[Amy's lab]*
Candidates: *{Amy, Penny, Sheldon}*
...
P0: *Hey, boyfriend.*
P1: *Can't talk. Spitball. Probably gonna die.*
Answer: P1 → *Sheldon*
Rationale: (Sheldon is Amy's boyfriend. After identify P0 is Amy, based on the relationship between Amy and Sheldon, P1 can be identified as Sheldon.)

- Status

Background: (from TBBT) *[The pub]*
P0: *So when do you guys plan on getting married?*
P1: *Uh, we're not sure. But I want to wait long enough to prove to my mother I'm not pregnant.*
P2: *May I have one of your fries?*
P1: *Of course. Can I have a bite of your burger?*
P2: *Absolutely not.*
P3: *Some perfect couple. He won't even share his food with her.*
Answer:P3 → *Leonard*
Rationale: (The aforementioned failure to determine Leonard's marriage led him to ridicule couples in harmonious relationships.)

Inside-Scene

- Background

Background: (from TBBT) *[Penny's apartment]*
Candidates: *{Amy, Penny}*
Bernadette: *Nah, you got this. Let's go for a drink.* ***I'll call Amy****.*
P0: *Okay, good. She seemed like she really wanted to go out tonight.*
P1 (phone ringing, running down stairs from outside penny's door): *Hey, girl.*
Answer: P1 → *Amy*
Rationale: (Bernadette said she will call Amy and P1 is the person who answers the phone.)

- Mention

Background: (from TBBT) *[The apartment]*
Candidates: *{Raj, Leonard, Sheldon, Amy}*
P0: *Mmm, I love how they put a waterfall at centre field. It really ties the whole stadium together.*
P1: *This is fun, huh? We get to see our friend throw out the first pitch, have a hot dog, watch the game.*
P2: *Whoa. Nobody said anything about watching the game.*
P3: ***Sheldon****, what did you expect?*
Answer: P2 → *Sheldon*
Rationale: (P3 mentioned the name of the person being questioned which is "Sheldon")

Evidence Type		Description
Linguistic style		Linguistic style refers to a character's individualized speech pattern. It consists of a selection of linguistic features such as vocabulary, syntactic patterns, rhythm, and tone. It also includes the use of elements such as direct or indirect, metaphor and irony.
Personality		Personality is a person's stable attitude toward objective facts and the habitual way of behavior that is compatible with it. We adopt a wider definition of personal traits as in (Li et al., 2020).
Fact	Attributes	Fact of a character's attributes in the TV series setting, such as race, profession, education level etc.
	Relations	A character's relationship with others that truly exist in the TV series setting, including both social relations and drama role relations.
	Status	Facts of a character's temporal emotional or psychological status in the time period when the scene happens.
Memory		The episodic memory about history events a character has in the previous show scenes. This also includes a rare case of a knowledge fact (i.e. the semantic memory) a character acquires from history scenes, which cannot be inferred from the facts of the character.
Inside-scene	Background	The character's identity can be inferred from the background introduction of scene, or from the description of the other characters' words.
	Mention	The character's name or alias is called by the other people.
Exclusion		The character's identity can be determined from the presence of characters in the scene and the other resolved characters.

Figure 6: The definitions of evidence types.

Exclusion

Background: (from `Friends`) *[Scene: Outside the Janitor's Closet, there are people having s*x and Mr. Geller is trying to give them some pamphlets.]*
Candidates: *{Monica, Chandler}*
Mr. Geller: *Kids, I spoke to a doctor and picked up this pamphlets on how to get pregnant. (He slides them under the door.*
P0: *(walking by with Chandler.) Hey dad!*
P1: *Hey.*
Mr. Geller: *(pause) Sorry to bother you again, but could you pass my pamphlets back? (They do so.) Thank you.*
Answer: P1 → *Chandler*
Rationale: (Monica is Mr. Geller's daughter. P0 called Mr. Geller dad so she is Monica. There are only two candidate so the other one is Chandler)

C Extended Study of Required Reasoning Types on our TVSHOWGUESS

This section provides an in-depth analysis of the types of reasoning used to infer evidence when guessing characters.

C.1 Our Annotation Schema of Reasoning Types

We define the following reasoning types with examples provided. A summary of our annotation schema of reasoning types can be found in Figure 7.

Multi-hop on Characters Reasoning on the basis of other characters that have already been guessed. Using the already guessed character as a bridge, users can employ history event or the relationship between characters to make guesses about the target character.The difference between multi-hop character and exclusion is that after identifying the other characters, the exclusion technique relies only on the list of characters provided for guessing, however, multi-hop character reasoning requires additional evidence such as relationship to infer the target character.

Background: (from `TBBT`) *[Angels Stadium]*
Candidates: *{Raj, Leonard, Sheldon, Amy}*
P5: *Hey, I hear you're a dermatologist.*
Emily: *Uh, yeah, I'm a resident at Huntington Hospital.*
...
P5: *I have some odd freckles on my buttocks. Can I make an appointment for you to look at them?*
Emily: *Um, okay, I guess.*
P0: *I'm with him three years, nothing. She's with him two minutes, and he's taking his pants off.*
Answer: P0 → ***Amy***
Rationale: (Using P5 (Sheldon) as a bridge and the couple relationship between Amy and him, we can identify P0 is Amy.)

Multi-hop on Textual Evidence Some evidences are not directly presented in the scene but can be inferred from the descriptions of context and dialogues. Using the inferred evidences as bridges people can multihop over personality, or fact, or event inferred from the text to guess the characters.

Reasoning Type	**Description**
Default Conjunction	No single piece of evidence can solve the task, hence the conjunction among multiple pieces of evidence is required. This is the default reasoning type if there are multiple evidence types labeled but no other reasoning types are labeled.
Multihop-Character	Task needs to be solved with the guessing results of other characters, then using the target person relation to or memory about the guessed ones to make the answer, *i.e.*, multihop with guessed characters as bridges.
Multihop-Textual	Task needs to be solved with the persona/fact/event not directly described in the scene but can be inferred from the context, *i.e.*, multihop over persona/fact/event inferred from dialog and scene context.
Commonsense attributes/relations of concepts/events	Task requires additional commonsense knowledge of attributes of daily concepts or social events, or their relations like causal relations between events. Those refer to the specific types of commonsense covered in ConceptNet- or Atomic-style KBs.

Figure 7: The definitions of reasoning types.

Background: (from TBBT) *[The apartment]*
Candidates: *{Amy, Leonard, Raj, Howard', Penny, Sheldon}*
Bernadette: *I like your suit.*
P0: *Oh, thanks. Got a couple new outfits for work.*
P1: *How does it feel knowing your fiancée's job is to go out and flirt with doctors, looking like that, while you sit here, you know, looking like this?*
...
Answer: P0 → *Penny*
Rationale: (P0 has a new job can be inferred from the textual evidence "Got a couple new outfits for work". Plus we know that Penny has a new job, we can determine that P0 is Penny)

Commonsense of Concepts/Events Task requires additional commonsense knowledge of attributes of daily concepts or social events, or their relations including causal/effect relations between an event and a social state or social relation. We restrict this category to be the aforementioned commonsense knowledge types, to distinguish from other relatively under-studied commonsense knowledge, such as the commonsense of dialogue flow required to work with our inside-scene evidence defined in Figure 6.

Background: (from TBBT) *[Capital Comics]*
Candidates: *{Howard, Sheldon}*
...
P0: *I know that if I had a wife or a fiancée, I'd ask her first before I invested money in a comic book store.*
P1: *He's right.*
Answer: P1 → *Howard*
Rationale: (A married or engaged person will answer "He's right". Howard is married.)

Default Conjunction A single piece of evidence will not solve this task; a combination between multiple pieces of evidence is needed to identify the person.

C.2 Analysis of the Human Annotation

Correlation between the Human Annotated Schema Categories Figure 2 visualizes the flow between (a) evidence types and the dependency of history and (b) evidence types and the reasoning

Reasoning Type	Friends(%)	TBBT(%)
Default	16.56	28.48
Multihop(Character)	3.97	13.91
Multihop(Textual)	5.30	5.30
Commonsense	4.64	0.66
No Complex Reasoning	69.54	51.66

Table 7: Percentage of the required reasoning types in the two TV shows, Friends and The Big Bang Theory.

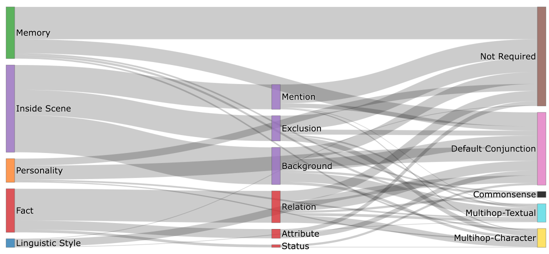


Figure 8: Visualization of the flow from the required evidence types to their required reasoning types.

types. Most evidence types correlate with history dependency. Personality and history dependency are most closely related. Default conjunction is the reasoning type that accounts for the largest percentage.

C.3 Experiments: Performance Decomposition on the Reasoning Types

We further studied the impact of the required reasoning types on the performance (the right column in Figure 9). In general there is a clear gap (on average ∼10%) between cases that require complex reasoning with those do not. The *Multihop-Textual* type is most challenging, because it requires both deep understanding of what the texts implies and multihop reasoning. There is not a clear performance difference between *Multihop-Character* and *Default Conjunction*, though the former is conceptually harder. We hypothesize this is because both

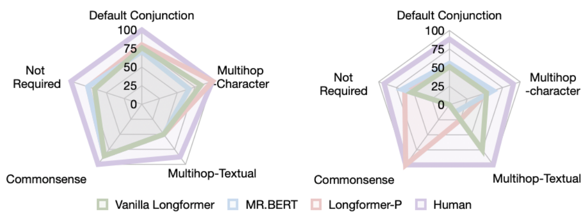


Figure 9: Performance breakdown according to our reasoning schema (left: Friends, right: The Big Bang Theory).

#Unsolvable		#Human Mistakes	
TBBT	Friends	TBBT	Friends
4882	2500	4921	
4895		4894	
4907		4910	
4908			

Table 8: Human Errors

types are beyond the reasoning ability of the model so the predictions largely rely on fuzzy matching of evidence – recall that we predict identities of main characters, so there can be a statistical bias of their context co-occurrence. The results on the *Commonsense* type fluctuate due to the relatively smaller ratio.

D Interface for the Human Study

Figure 10 shows the interfaces of the human study.

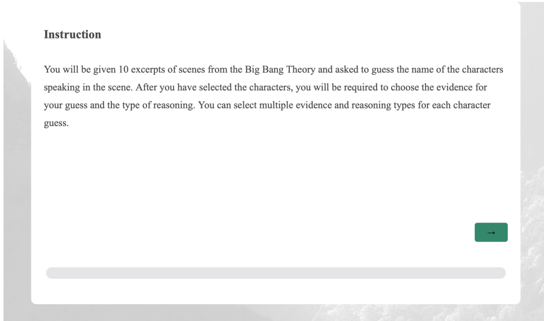
E Examples of Human Errors

Table 9 provides an example of unsolvable cases and Table 10 provides an example of human mistakes. The human mislabeled characters are marked as red.

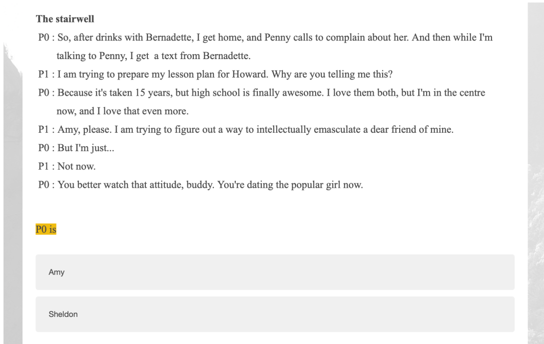
We further provide all the scene IDs on which our human tester makes incorrect predictions in Table 8.

F Details of Human Study and Discussions on the Challenges of History Retrieval

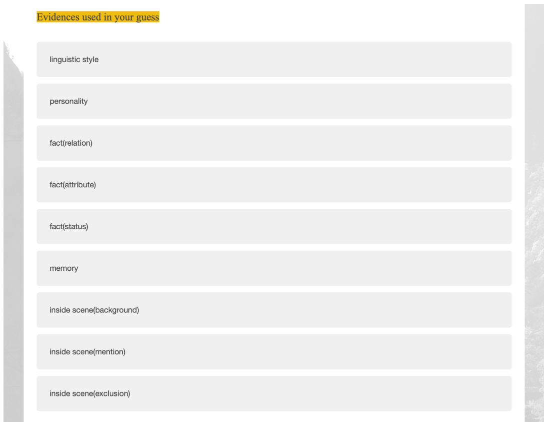
Our experiments show that the history dependency challenges existing models. Finding the evidence history scenes for such cases is essentially a retrieval task (but without groundtruth). To see how it brings new challenges to existing semantic search, we applied a state-of-the-art model to retrieve the history scenes and conducted an additional human study to evaluate the results.



(a) Introduction page of human study.



(b) Task 1: character guessing task



(c) Task 2:identifying used evidence types.

(d) Task 3: identifying used reasoning types .

Figure 10: interfaces of human studies.

Task We conduct the study on scenes in our human annotation sets that have the *Memory* type labeled. With each scene as a query, we retrieve from a window of 20 previous scenes with a state-of-the-art model[8] The window size is decided so as to guar-

[8]We use the `all-mpnet-base-v2` model from https://sbert.net/ that reports the top-1 performance on 14 sentence embedding tasks and 6 semantic search tasks.

Unsolvable Case
08x02 4882 **Background:** (from `TBBT`) *[the Apartment]* **Candidates:** *{Howard, Sheldon, Raj, Amy, Leonard, Penny}* P0 : I recently read that during World War Two, Joseph Stalin had a research program to create supersoldiers by having women impregnated by gorillas. P1 : What a sick use of science. P2 : Hey, as long as the baby's healthy. P3 : I wonder if Stalin considered any other animals. P4 : Hippos are the deadliest creature. A half-human, half-hippo soldier would be pretty badass. P1 : Yes, but when they're hungry-hungry, you can stop them with marbles. P0 : Yeah, the correct animal for interspecies supersolider is koala. You would wind up with an army so cute it couldn't be attacked. P2 : But half-man, half-owl could fly... P0 : The answer is cuddly soldiers with big flat noses. Moving on. P1 : So, Penny, when's the new job start? P5 : Next Monday. Bernadette : Did you get a chance to look over the materials I gave you? P5 : Uh, not yet, but I will. Bernadette : Great. When? P5 : I said I'll get to it. P0 : I'm sensing awkwardness, am I right? P3 : Yes. P0 : Swish. Bernadette : I don't want to be pushy, but you've never done pharmaceutical sales before. It seems like you could use this time to get a head start. P5 : Well, the first few weeks will be all training. They'll tell me everything I need to know. Bernadette : But imagine how impressed they'd be if you showed up already familiar with the material. P5 : Okay, so what, you want me to be like a teacher's pet? Bernadette : Couldn't hurt. P4 : Mm, I don't know. Who here has ever been hurt because they were the teacher's pet? P0 : It was like the rest of the class wanted Ms. McDonald to forget the quiz. **Answer: P0: Sheldon, P1: Howard, P2: Raj, P3: Amy, P4: Leonard, P5: Penny**

Table 9: Example of unsolvable case.

Mistake
08x04 4921 **Background:** (from `TBBT`) *[Penny's partment]* **Candidates:** *{Raj, Penny}* P0 : I'm so glad we could work this all out. P1 : Yeah, me, too. Emily : You know, we should have dinner one night with you and Leonard. P1 : Oh, we would love that. P0 : Great. background : (both chuckle) P1 : Okay, good night, guys. Emily : All right, night. P1 : Bye. Emily and Penny (simultaneously) : I hate her. **Answer: P0: Raj, P1: Penny**

Table 10: Example of mistake.

antee that at least one required memory appears in the window, according to our human annotation process. The task of human study is to recognize whether the top-3 returned scenes contain at least one related history scene.

Results The same annotators working on the study in Section 4 are asked to evaluate the retrieved scenes. The results show that the recall of the top-3 results from this state-of-the-art model is very low (35.5%). We observe the following major reasons for this difficulty in scene retrieval: (1) the queries are scenes with structures, which leads to different query formats from standard IR tasks; (2) many relevant scenes are not similar to the query scenes in the semantic space, but is associated with the query in specific aspects or even forms analogy to the query scene; (3) some scenes require a multi-hop retrieval, especially when combined with ToM modeling (reasoning about what the others knows).

All these challenges are non-trivial, and calls for further studies on semantic search to address.

G Model Checklist

We implement our baselines based on HuggingFace Transformers.[9] We use the pre-trained `allenai/longformer-base-4096` and `bert-base-uncased` models. We train all the models with the Adam optimizer.

We train our model on a single A100 GPU. It takes around 1 hour and 40 minutes to train a Longformer-based model. It takes around 2 hour and 10 minutes to train a multi-row BERT model. For all the models, we train in total 40 epochs. But the models usually converge in less than 20 epochs.

Hyperparameters We set the number of rows in MR. BERT to 12, to maximize the usage of GPU memory. We set the maximum length of Longformer to 2000, which can handle the lengths of most of the input scenes. The window size is set to 256. We set the learning rate to 2e-5.

We report our result with a single run. However, for each model we run twice; and we found the average development accuracy varies less than 0.5%.

[9] https://github.com/huggingface/transformers