PASTA: A Dataset for Modeling PArticipant STAtes in Narratives

Sayontan Ghosh[♢] **Mahnaz Koupaee**[♢] **Isabella Chen**[♢†]
Francis Ferraro[♠] **Nathanael Chambers**[♣] **Niranjan Balasubramanian**[♢]
[♢]Stony Brook University, USA [♠]University of Maryland, Baltimore County, USA
[♣]United States Naval Academy, USA
{sagghosh, mkoupaee, niranjan}@cs.stonybrook.edu
isabellachenusa@gmail.com ferraro@umbc.edu nchamber@usna.edu

Abstract

The events in a narrative are understood as a coherent whole via the underlying states of their participants. Often, these participant states are not explicitly mentioned, instead left to be inferred by the reader. A model that understands narratives should likewise infer these implicit states, and even reason about the impact of changes to these states on the narrative. To facilitate this goal, we introduce a new crowdsourced English-language, *Par*ticipant *Sta*tes dataset, PASTA. This dataset contains inferable participant states; a counterfactual perturbation to each state; and the changes to the story that would be necessary if the counterfactual were true. We introduce three state-based reasoning tasks that test for the ability to infer when a state is entailed by a story, to revise a story conditioned on a counterfactual state, and to explain the most likely state change given a revised story. Experiments show that today's LLMs can reason about states to some degree, but there is large room for improvement, especially in problems requiring access and ability to reason with diverse types of knowledge (e.g., physical, numerical, factual).[1]

1 Introduction

Understanding narrative text requires forming a coherent representation of the scenario, including filling in details that are unstated in the text. One type of detail that is usually not mentioned is the state of its participants[2] (e.g., ''she unlocked the door'' implies the possession state that ''she has a key''). The reader easily infers these *implicit* states and their causal relationships with the narrative's *explicit* events, creating a detailed mental picture of the described world that is only partially observable from the text. Many cognitive theories have been proposed to capture aspects of this in their representations, such as scripts (Schank and Abelson, 1975), frames (Fillmore, 1985), and state/time formalisms (Galton, 1990). Without committing to any one particular formal theory, this paper adds a theory-agnostic resource to test such theories by listing implicitly assumed participant states in simple narratives.

Consider the story in Figure 1 from the ROCStories corpus (Mostafazadeh et al., 2016). Humans create a detailed mental representation of this spilled-soda scenario by inferring its commonsense states. In this story, using our commonsense knowledge about emotions and habituals, we can infer from the first two lines that Kate's mother liked keeping her car clean (a state about *Kate's mother*). Similarly, based on our physical commonsense of a lid, i.e., that lids prevent spilling, we can also assert from the spill that the soda's lid was loose (a state about the *soda*). We can also reason about the likely change to the story due to a counterfactual state, i.e., if the soda's lid was tight, then most likely the soda wouldn't spill. To the best of our knowledge, no such resource exists that captures this kind of participant state knowledge.

To capture this type of commonsense knowledge needed to understand and reason about participant states in narratives, we introduce PASTA, a crowd-sourced dataset in English. As shown in Figure 1, for a given story S, PASTA provides a participant state α that is likely to be inferred from S, a perturbation state α' that is counterfactual to S along with the minimal changes to S that are required to make α' likely

[†]Work done during internship at Stony Brook University.

[1]Code and the dataset are available at https://github.com/StonyBrookNLP/pasta.

[2]We define participants to include both animate entities and inanimate objects in the narratives.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1283–1300, 2023. https://doi.org/10.1162/tacl_a_00600
Action Editor: Xiaojun Wan. Submission batch: 1/2023; Revision batch: 5/2023; Published 11/2023.

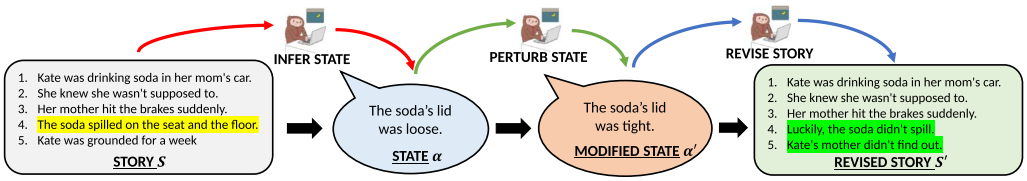

Figure 1: For a given story S, **PASTA** provides an (unstated) inferred state α, a minimal set of justification sentences (yellow highlight), a counterfactual state α', and a revised story S', such that α' can be inferred from it.

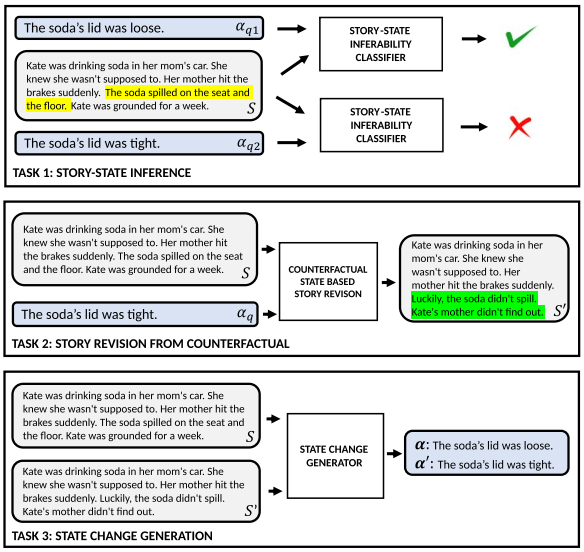

Figure 2: Examples of three PASTA tasks. Input for each is on the left. The boxes indicate systems required to solve the tasks with example output on the right.

to be inferred from the revised story S'. PASTA includes 10,743 instances of these story/state/counterfactual/revision tuples. With this new dataset, we hope to enable models to make the kinds of state-based inferences that move beyond surface text understanding and lead to deeper reasoning. To this end, we describe three new state-based reasoning challenges with PASTA which are illustrated in Figure 2.

The first is **Story State Inference**: Given a story and an inferred participant state, predict if the state is likely to be inferred from a given set of sentences in the context of the story. We formulate this as a binary classification task, and we create contrastive examples for training and evaluation purposes to guard against artifact-based reasoning. This can be seen as a form of textual entailment, a capability useful for applications such as question answering (Harabagiu and Hickl, 2006; Trivedi et al., 2019), claim verification (Yin and Roth, 2018; Hanselowski et al., 2018), etc.

The other two challenge tasks are generative. The second task, **Story Revision for Counterfactual States**, measures the ability to reason about counterfactuals. Given a story and a counterfactual state (i.e., a state that is not consistent with the story), the task is to revise the story such that the counterfactual state is now likely to be inferred from it. These types of counterfactual revisions serve as a test of reasoning (Qin et al., 2019) and can support interactive story generation tasks (Goldfarb-Tarrant et al., 2019; Brahman et al., 2020). The third task, **State Change Generation**, requires the model to take a story and its perturbed version as input and then generate the two corresponding states (e.g., 'lid was loose' and 'lid was tight') that explain the differences in the way they unfold. From an application perspective, generating the underlying states that account for the differences between two narratives can assist with fake news detection using reliable sources (Figueira and Oliveira, 2017; da Silva et al., 2019; Ghadiri et al., 2022) and information fact checking (Brandtzaeg et al., 2018).

These three challenge tasks require a unique combination of commonsense abilities, thus helping to evaluate models on reasoning and knowledge capacity. These tasks require not only basic entailment ability, but also knowledge (numerical, factual, physical, etc.) and broader narrative understanding. Having just one of these abilities will not suffice. To evaluate current models for these capabilities, we benchmark the LLMs T5 (Raffel et al., 2020), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT3 (Brown et al., 2020). For the generative tasks, we evaluate model performance through extensive human and automatic evaluations. The results show that, though these models can reason about states to some degree, there is substantial room for improvement on all tasks, suggesting avenues for future research.

2 Related Work

There are many formal theories on mental states and reasoning. The seminal work by Schank and Abelson (1975) introduced scripts as a way to structure knowledge about stereotypical event sequences with their participants. Frames (Fillmore, 1985) and theories of time (Galton, 1990) provide related views. This paper does not commit to a formal theory, instead providing a challenge dataset to test aspects of them. Statistical work on events (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Balasubramanian et al., 2013; Ferraro and Van Durme, 2016; Sha et al., 2016) curated event knowledge in an unsupervised manner from large text corpora. This paper augments their view with *state-based* knowledge about event participants.

More recent work by Speer et al. (2017), Sap et al. (2019), and Hwang et al. (2021) capture everyday inferential knowledge associated with an action performed by someone. This knowledge is organized through a fixed set of relationship classes between action and inferences. A subset of these classes is about participant mental states, but this commonsense knowledge is non-contextual in nature. In contrast, our work requires inferring commonsense knowledge about participant states in the context of coherent narratives.

Most similar to our work is the TIME-TRAVEL dataset by Qin et al. (2019). It includes the Counterfactual Story Rewriting task to edit a short story based on a counterfactual context. The authors insert an explicit counterfactual at a fixed position (2^{nd} sentence) in the story, and the revision task is then conditioned on this observed change. It is a language modeling generation task. In contrast, our work introduces *unobserved* counterfactual outside of the story's text, and the revised story must be generated with deeper state-based reasoning. This introduces additional complexity for the revision task. Also, TIME-TRAVEL requires the revisions to be restricted to the story ending, which cannot be assumed in our setting. Our states can be inferred from any part of the story.

Bhagavatula et al. (2019) proposed tasks that predict a plausible hypothesis for two given observations, and curate a dataset for the same. Their work mainly focuses on *what happened in-between?* type of inferences. Mostafazadeh et al. (2020) introduced the GLUCOSE dataset, which focuses on several types of causal knowledge that are required to explain a causal event in narrative text. Neither of these focuses entirely on implicit states (some GLUCOSE annotations are relevant, but not directly so), and neither addresses story revision in the face of counterfactual changes.

Recent work on understanding entity states has mostly focused on tracking entity state change in text. Dalvi et al. (2018) introduced PROPARA, which captures physical state changes (creation, destruction, and movement), Bosselut et al. (2018) proposed the task of tracking ingredients in cooking recipes, and Rashkin et al. (2018) tracks the emotional reactions and motivations of characters in simple stories, for a fixed/small set of attributes.

Tandon et al. (2019) introduced the WIQA dataset for analyzing the effect of perturbing a process described by a procedural text on the elements (entities, events, etc.) of the text, as an influence graph of the process. However, the influence graph was assumed to have a fixed causal structure. It captured a very limited set of cause-effect relationships obtained as a result of analyzing perturbations that either accelerated or decelerated the main outcome of the process. Tandon et al. (2020) introduced a dataset for tracking state changes in procedural text as a set of state change tuples of *entity, attribute, before-state*, and *after-state* for each step of the process. The elements of the tuples were in free-form text instead of belonging to a set of pre-defined categories.

Our work differs from the above in several key ways: (i) participant states are unstated, (ii) participant state inferences do not depend on sentence ordering assumptions, (iii) state perturbations affect the entire discourse of the narrative, and (iv) captures how the participant states change between an original and a perturbed narrative.

3 PASTA: PArticipant STAtes

PASTA is a dataset of story pairs (S, S') where each story S has a revised version of itself, S', that hinges on a particular state that was changed in its revision. The story pairs thus have corresponding state pairs (α, α'), containing an original state α and its counterfactual α'. Refer to Figure 1. These story/state pairs allow us to analyze unique narrative challenges. We can test if a model can identify whether a given participant state is consistent with it. We can ask what would happen if

an assumed story state is no longer true. We can also ask if a model can identify what/how a state changes between two similar but different stories. This section describes PASTA, the crowd-sourcing process that created it, quality control details, and its basic statistics.

3.1 Data Annotation

To create the PASTA dataset, we use stories from the extended ROCStories (Mostafazadeh et al., 2016) corpus for annotation by crowd workers. ROCStories narratives describe a rich set of causal and temporal commonsense relations between daily events, and its stories are short enough that the world described by them are self-contained. They thus are a good fit for testing state inferences.

Figure 1 illustrates the process followed to collect responses from the crowd workers.[3]

The annotation process has four main steps:

1. **Infer a participant state**: For a story S, the annotator infers a participant (or object) state α that is likely to be true at some point in S; α is a free-form sentence. Most stories have several inferrable states, so the annotator may identify whatever jumps out to them the most.

2. **Select minimal justification sentences:** For the inferred α, the annotator selects the minimal set of sentences J_α^S in S, that they used to infer α from S.

3. **Perturb the state:** The annotator perturbs α to create α' such that α' is very unlikely to be true for the story S. α' is also a free-form sentence.

4. **Revise the story:** The annotator revises S into S', so that α' can be inferred from S' but α is unlikely to be inferred from S'. The annotator is instructed to make minimal revisions in order to avoid creating S' with other narrative side-effects.

We provide detailed instructions about how to infer a state, and these are repeated not just in the instructions and examples, but also in the actual form the participants fill out. The inferred state must be a property or attribute of a participant or object (e.g., *she was angry* or *the rock is heavy*); it must not be an action (e.g., *Susan is running* or *Jake cooks food*); and it must not be explicitly stated in the story. These constraints ensure that the states are not readily available from the story text, and must be inferred by reasoning and world knowledge. The next section describes how we monitored the workers and mitigated improper responses.

3.2 Quality Control

For crowdsourcing the data collection, we used the Amazon-MTurk platform (AMT). Each story was provided to three different crowd workers for annotation. We priced the HIT at $\$0.35$ based on initial worker response times and interest gleaned from multiple pilot runs. For filtering out noisy data from the collected responses, we follow a two-stage filtering process.

Stage 1: We only allowed workers with a long history of consistent performance who satisfied the following criteria:

1. have responded to at least 5000 HITs
2. have at least 98% accuracy on their past HITs
3. must reside in USA or Canada; this helps to prevent language-based artifacts

Although the above is strict, we still observed responses that did not follow the instructions. One difficulty was how workers wrote their revised stories. Even minor changes to the original story can render it logically inconsistent, so care is needed to ensure the counterfactual is inferrable while still maintaining coherence. Other annotation errors were 'states' describing actions, states directly mentioned in the story, and non-entailed states.

Stage 2: Despite the above errors, we received excellent responses with clear states and interesting revised stories. This gave us confidence that the task is achievable, but it just needed expert crowd workers. To this end, we performed an ''expert review'' of the responses to identify ''proficient workers'': workers who can perform the task with a high degree of correctness. Our expert reviewers are two student researchers who work in the field of common-sense reasoning and NLP in general. Stage 1 resulted in a total of 9656 responses from 136 workers. The experts

[3]The project was reviewed and approved by the local institutional review board for human subjects research.

evaluated a subset of these to identify proficient workers by using the process described below:

1. For each worker, we manually evaluated their performance on a random sample of their responses.
2. The number of evaluated responses for each worker was decided by the formula below. If the i^{th} worker provides n_i responses, then the minimum number of their responses, e_i, that needs to be expert-reviewed to evaluate their proficiency is given by:

$$e_i = \begin{cases} 0.3 * n_i & n_i < 100 \\ 0.2 * n_i + 10 & 100 \leq n_i < 150 \\ 40 & 150 \leq n_i \end{cases}$$

3. Each evaluated response was categorized as correct or reject. A response was rejected if there was an error in any of the four steps of the annotation process. A response is correct if all the components of the annotation adheres to the instructions.
4. A worker was identified as proficient if they submitted ≥ 50 responses with a rejection rate $\leq 20\%$. After identifying proficient workers, all other responses from proficient workers were then auto-accepted. We also kept the smaller number of non-reject responses that our experts labeled from non-proficient workers.

With this process, we identified 28 workers who were proficient. We accepted all of their annotations, totaling $\sim 6,000$. To this we added the annotations the experts accepted in the review, which added another 360 high-quality instances. We then ran a second round of data collection using only the proficient workers. We added this to the high-quality instances from the first round to form our full PASTA dataset.

The responses in the pool of expert-reviewed responses were used to create the test set of the data. We also made sure that there is no story overlap in the train, validation, and test sets.

3.3 Dataset Statistics

PASTA includes a total of 10,743 (8476 train, 1350 validation, and 917 test) 4-tuples. Each 4-tuple is a story S, an associated inferred state α, counterfactual state α', and a revised story S'. Annotators almost always changed the justification sentences of the inferred state in order to revise the story. Instructions to make minimal changes to the revised story results in a high degree of similarity between the original and revised stories. On average 1.5 out of 5 story sentences are changed to create the revised story, with 90.3% average token overlap between them. Similarly, the inferred state and its counterfactual on average show high lexical similarity with 72% token overlap, both having similar token length. Additional statistics can be seen in Table 1.

# of unique stories	5,028
Avg. # of tokens in an inferred state	5.7 tokens
Avg. # of tokens in a perturbed state	6 tokens
Avg.# of justification sentences for a state	1.5
Avg. # of sentences revised in a story	1.48
% of justification sentences that are revised	90.54%
% of revised sentences that were justification	91.9%
% tokens in inferred state, common in perturbed state	71.9%
% story tokens common in revised story	90.3%

Table 1: PASTA dataset statistics.

4 State-based Reasoning Tasks

Inferring each component of a PASTA 4-tuple requires a different commonsense reasoning ability about a participant's state in a narrative, which enables us to use PASTA to test models for these abilities. As illustrated in Figure 2, we introduce three PASTA tasks, one classification and two generative, each of which can be used to evaluate current NLP models for the capabilities required to understand a participant's state in a narrative text. In the subsections below we provide the motivation and formal task definition for each task.

4.1 Story State Inference

We propose a classification task to evaluate a model's ability to understand what state is likely or unlikely to be inferred from a story. We deem a state is *likely to be inferred* from a story if a typical human reading the story would conclude that the state is most likely true. To test this capability in models, we pose the Story State Inference classification task.

Task Definition: Given a story S, a 'query' state α_q, and a supporting set $\mathbf{s}$, which is a subset of the sentences in S, the task is to predict whether α_q is likely to be inferred from $\mathbf{s}$ in the context of S.

Effects of Data Collection on Performance: We provide additional dataset analysis in subsection 6.1 to analyze the robustness of our data collection procedure that helped avoid unintended artifacts in the data for this task.

4.2 Story Revision for Counterfactual States

A model that can understand participant states in narrative text should also be able to reason about counterfactual states and their potential effects on the narrative. We introduce the Story Revision for Counterfactual States task to address this.

Task Definition: Given a story S, and a participant state α_q that is counterfactual to S (a state that is not consistent with S), make minimal revisions to S to generate S' such that α_q is unstated in S' and can be inferred from S', i.e., $P(\alpha_q|S') \approx 1$ and $P(\alpha_q|S') \gg P(\alpha_q|S)$.

4.3 State Change Generation

A corollary of being able to reason about the effects of a counterfactual state on the discourse of a narrative is the ability to identify the state changes (and how they changed) which led to the new narrative. In other words, when given a revised story with its original, what original state and its counterfactual explains the change? To assess this, we introduce the State Change Generation task.

Task Definition: Given a story S and its revision S', the task is to generate participant states α, α' that describe the change of state from S to S', i.e., $P(\alpha|S) \gg P(\alpha|S')$ and $P(\alpha'|S') \gg P(\alpha'|S)$.

4.4 Task-specific Data Creation

The three tasks above use the PASTA 4-tuple (S, α, α', S') to create task specific data instances in the following manner:

1. Story State Inference: Let $S = (s_1, \cdots, s_5)$ and $S' = (s'_1, \cdots, s'_5)$. We create four data instances for the task, positive data instances $((S, \mathbf{s}, \alpha), 1)$ and $((S', \mathbf{s}', \alpha'), 1)$, and negative instances $((S, \mathbf{s}, \alpha'), 0)$ and $((S', \mathbf{s}', \alpha), 0)$. The supporting set $\mathbf{s}$ for S is J^S_α, i.e., the minimal set of sentences used to infer α from S. For S', $\mathbf{s}' = \{s \in \{s'_1, \cdots, s'_5\} | s'_i \neq s_i, \forall\, i \in 1 \text{ to } 5\}$, i.e., the set of sentences in S that were changed when revising S to S'.

2. Story Revision for Counterfactual States: We created two data instances for the task of the form $((S, \alpha'), S')$ and $((S', \alpha), S)$.

3. State Change Generation: We created two data instances for the task of the form $((S, S'), (\alpha, \alpha'))$ and $((S', S), (\alpha', \alpha))$.

5 Experimental Setup

To establish modern baselines and measure their performance, we built benchmark models from GPT3, T5, BERT, and RoBERTa. This section describes how each was setup for the three tasks.

5.1 GPT3

We benchmarked GPT3 with few-shot prompting (Brown et al., 2020) on the two generation tasks (Story Revision and State Change). We created prompts with task examples from the training set, followed by an incomplete prompt from the eval set that the model must complete. For the Story Revision for Counterfactual States task, the prompt included n examples followed by the final query: $(S_1, \alpha'_1, S'_1) \cdots (S_n, \alpha'_n, S'_n)$ $(S_q, \alpha'_q, -)$ where (S_i, α'_i, S'_i) is the i^{th} task example. The model must generate S'_q for the final query $(S_q, \alpha'_q, -)$. Similarly, the State Change Generation task uses a similar prompt: $(S_1, S'_1, \alpha_1, \alpha'_1) \cdots (S_n, S'_n, \alpha_n, \alpha'_n)(S_q, S'_q, -, -)$.

To select prompt examples, we tried three approaches. **(i) EXPERT CURATED:** We selected a fixed set of diverse, unambiguous examples that requires multi-step reasoning and covers different type of states, and used the same prompt examples for all the query instances, **(ii) RANDOM SELECTION:** We randomly selected examples, **(iii) NEAREST NEIGHBOR** (Liu et al., 2022): For each query instance, we selected examples that were most similar to it. For this, we computed the cosine similarity between the [CLS] representation of the instances obtained from RoBERTa-large fine-tuned on the Story State Inference task. For each approach, we tried creating prompts with 5, 10, and 15 examples. Prompt examples were selected from a set of 200 high-quality, expert-selected instances drawn from the training set, similar to West et al. (2022).

We treat the number of prompt examples and their selection as hyperparameter combinations, and evaluated each of them on 200 random samples from the validation set. Since human evaluations are expensive, we use BERTScore, which

	# of examples in prompt		
APPROACH	**5**	**10**	**15**
EXPERT CURATED	81.6	81.6	82.5
RANDOM SELECTION	81.2	81.8	82.3
NEAREST NEIGHBOR	81.7	83.3	82.0

(a) Story Revision for Counterfactual States

	# of examples in prompt		
APPROACH	**5**	**10**	**15**
EXPERT CURATED	53.0	52.8	51.9
RANDOM SELECTION	51.4	51.2	52.1
NEAREST NEIGHBOR	52.1	51.5	50.2

(b) State Change Generation

Table 2: GPT3 hyperparameter selection. Few-shot performance (BERTscore) of GPT-3 for combinations of (i) prompt example selection approach and (ii) # of examples in prompt.

has the highest correlation with human-evaluated validity of output, among the automatic metrics we tried (see Table 9). Table 2 shows that the combinations perform roughly similarly but there is a two-point gap between the best and the worst combination. For the Story Revision for Counterfactual States task, we use NEAREST NEIGHBOR with 10 prompt examples, and for the State Change Generation task, we use EXPERT CURATED with 5 examples.

We used the `text-davinci-002` GPT3 model for both tasks. We set the generation temperature parameter to 0.9, frequency penalty to 0.5, and maximum generation length to 100.

5.2 T5

We benchmarked base (T5-b) and large (T5-l) variants of T5 on all three state-based tasks by fine-tuning them on the task-specific instances created from PASTA, as explained in Section 4.4. Examples of T5 input-output format for each task are shown in Figure 3. For all the tasks, T5-b and T5-l were trained for 7 and 5 epochs, respectively. For model training, we used the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 10^{-4} and weight decay of 10^{-6}. For T5-l the batch-size for tasks 1, 2, and 3 were 8, 4, and 4, respectively. Whereas for T5-b, the corresponding batch-sizes were 16, 12, and 10. For text generation, we used nucleus sampling with 0.93 top-p; 100 as max generation length.

Task 1. **Story State Inference**

Input: infer_state story: I live in the north. We usually have snow here for Christmas. This year is different. * It's 50 degrees out! * We will have a green holiday this year.
state: I live in the North Pole.</s>

Target Output: false</s>

Task 2. **Story Revision for Counterfactual States**

Input: revise story <extra_id_1>: Connor had a very busy workload that day. <extra_id_2>: He forgot to eat breakfast and grabbed a dozen doughnuts on the way. <extra_id_3>: People were greeting him and taking one doughnut each time. <extra_id_4>: When he arrived at his desk, there were no doughnuts left. <extra_id_5>: He went on with his busy day on an empty stomach.
state: Connor only has six coworkers.</s>

Target Output: <extra_id_4>: When he arrived at his desk, half the donuts were gone. <extra_id_5>: He ate a donut before starting work.</s>

Task 3. **State Change Generation**

Input: change story1: Joe was stopping at a drive thru for breakfast. He ordered a cup of coffee and breakfast sandwich. Joe was trying to eat and drive at the same time. Joe accidently dropped his coffee in his lap. Joe had to go home and change his pants.
story2: Joe was stopping at a drive thru for breakfast. He ordered a cup of coffee and breakfast sandwich. Joe was trying to eat and drive at the same time. Joe accidently dropped his coffee in his lap. Joe had to go to the hospital. </s>

Target Output: state1: The coffee wasn't hot enough to seriously burn Joe. **state2:** The coffee was hot enough to seriously burn Joe.</s>

Figure 3: Examples of T5 formats. An asterisk $*$ is prepended to the supporting set sentences in the stories in task (1). The output in task (2) includes special tokens like <extra_id_4> and <extra_id_5> to indicate revisions to the 4^{th} and 5^{th} sentences.

5.3 BERT \ RoBERTa

We benchmarked base (BERT-b) and large (BERT-l) variants of BERT-uncased, and base (RoBERTa-b) and large (RoBERTa-l) of RoBERTa on only the Story State Inference task since they are non-generative models. The input format for the models is identical to that of T5. For all the models, we used the AdamW optimizer with a learning rate of $5e{-}6$ and weight decay of $1e{-}6$. The large and base models were trained for 5 and 7 epochs respectively.

T5-b, BERT-b, BERT-l, RoBERTa-b, and RoBERTa-l were trained on an NVIDIA-TITAN-X 24GB, and T5-l was trained on an NVIDIA-A6000 48GB GPU.

6 Results and Analysis

We now analyze the performance of these recent language models on the three PASTA tasks.

6.1 Story State Inference

We evaluated model performance with standard accuracy and contrastive accuracy. In contrastive

accuracy, the model gets a point only if it makes correct predictions for both inferred and the counterfactual states for a story. For all models, we train with five random seeds and report their average performance with standard deviation.

Human Evaluation: We conducted human evaluation on the task instances (see Section 4.4.1) created from PASTA. We randomly selected 200 4-tuples from the test set,[4] and created 800 story-state inference instances from them. Each task instance, $(S, \alpha_q, \mathbf{s})$ was evaluated by three crowd workers who rated the likelihood of inferring α_q from $\mathbf{s}$ in context of S, on a 5-point Likert scale - Extremely unlikely, Unlikely, Cannot Say, Likely, and Extremely likely. We threshold the Likert value to a binary 0/1 value with the mapping $\{$Extremely unlikely to Cannot Say$\} \rightarrow 0$ and rest $\rightarrow 1$. The human prediction for an instance was computed by majority voting, which, along with its true label, was used to compute the human performance.[5]

Story State Inference is a Hard Task Table 3 shows that even for just standard accuracy, there is room for improvement (7.8%) when comparing the best performing model (RoBERTa-l) to humans on this simple binary classification task. This performance gap further increases to 10.5% when considering contrastive accuracy. Increasing model size from base to large yields 3.7% (BERT) to 8% (RoBERTa) gains on standard accuracy. For contrastive measure, both base and large variants of each models fare substantially worse, with performance drops ranging from 5.4% (RoBERTa-l) to 9.8% (BERT-b). For humans, the corresponding performance drop is only $\sim$ 2.7%. This suggests that predicting whether a state is likely to be inferred from a story is difficult for these LLMs, even when fine-tuning on a relatively large number of examples.

We also analyze the performance of the models when they don't have direct access to the justification sentence information in the story. We fine-tuned large variants of the three baseline models on this task. From Tables 3 to 4, we see that the task performance drops across all the models on both evaluation metrics, with a 2.4% to 3.5% drop in accuracy, and 3.3% to 5.5% in contrastive accuracy. The gap between the human performance and the best performing model is still substantial. This shows that justification sentences are indeed important to solve the task, but the models often still make reasonable decisions without them.

	Accuracy (%)	Contrastive Accuracy (%)
BERT-b	73.8 ± 0.3	64.0 ± 0.5
T5-b	79.8 ± 0.6	70.7 ± 0.5
RoBERTa-b	81.2 ± 0.6	73.0 ± 0.8
BERT-l	77.5 ± 0.4	68.7 ± 0.7
T5-l	83.1 ± 0.9	75.3 ± 1.4
RoBERTa-l	89.1 ± 0.4	83.7 ± 0.5
Human*	96.9	94.2

Table 3: Story State Inference - Model evaluation: Accuracy is the % of instances where the model made correct predictions. Contrastive Accuracy gives a credit to the model if it correctly predicts the inferability for both the inferred and counterfactual states for a story.

	Accuracy (%)	Contrastive Accuracy (%)
BERT-l	74.9 ± 0.3	64.6 ± 0.3
T5-l	79.6 ± 0.6	69.8 ± 1.0
RoBERTa-l	86.7 ± 0.4	80.4 ± 0.6
Human*	93.5	88.9

Table 4: Story State Inference - without the justification sentences: Model performance on a harder variant of the Story State Inference task, where they don't have direct access to the justification sentences of a state when predicting it's inferability for a story.

Importance of the Data Collection Design: It is important to note that we included contrastive examples in our train-set. To illustrate its importance, we trained a model on the dataset created from just the original stories ($\mathcal{D}_1 = \{((S_i, \mathbf{s_i}, \alpha_i), 1), ((S_i, \mathbf{s_i}, \alpha_i'), 0)\}_{i=1}^{N}$), and another on the modified stories ($\mathcal{D}_2 = \{((S_i', \mathbf{s_i'}, \alpha_i'), 1), ((S_i', \mathbf{s_i'}, \alpha_i), 0)\}_{i=1}^{N}$). We then trained and tested on these different dataset, results of which are reported in Table 5.

Generalization accuracy is significantly worse if we had only constructed positive and negative states for a collection of stories. For example, training on $\mathcal{D}_1^{tr}$ and testing on $\mathcal{D}_2^{te}$ leads to an 11.1% drop in accuracy compared to in-distribution test on $\mathcal{D}_1^{te}$. For $\mathcal{D}_2^{tr}$, the corresponding drop is 6.6%, which supports the quality

[4] Model performance for this test subset differed by $< 0.5\%$ from that of the overall test set

[5] The instance label assignment is explained in Section 4.4.

Train data	Test data		
	$\mathcal{D}_1^{te}$	$\mathcal{D}_2^{te}$	$\mathcal{D}_1^{te} \cup \mathcal{D}_2^{te}$
$\mathcal{D}_1^{tr}$	90.2(84.4)	79.1(70.8)	84.8(77.9)
$\mathcal{D}_2^{tr}$	81.5(73.6)	88.1(82.6)	84.8(78.1)
$\mathcal{D}_1^{tr} \cup \mathcal{D}_2^{tr}$	90.2(85.3)	88.1(82.3)	89.1(83.7)

Table 5: Story State Inference - Dataset Analysis: Accuracies and contrastive accuracies of RoBERTa-large when trained and tested on dataset partitions created from original stories ($\mathcal{D}_1$), modified stories ($\mathcal{D}_2$), and their union ($\mathcal{D}_1 \cup \mathcal{D}_2$). tr and te denotes the corresponding training and test splits.

	Acceptability			
	% Inferable (A)	% Logical (B)	% ALL (A & B)	Minimal Revision
GPT3 - FS	50	86	48.5	86.33
T5-b FT	41.0	77.0	34.0	91.39
T5-l FT	58.5	84.0	54.0	89.17

Table 6: Story Revision for Counterfactual States - Human evaluation: % of model generated stories that satisfy evaluation criteria. **Inferable** and **Logical** assure coherence of the revisions with the required entity state. **ALL** indicates generations that satisfied both the criteria. **FS** means few-shot learning, **FT** means finetuned on PASTA dataset.

of our stories/states and shows that both original and counterfactual state inferences are learnable. Had we not collected the revised story, then models could potentially learn artifact-based heuristics (e.g., guessing whether the state is original or modified) resulting in the lack of generalization that we observe here. Because PASTA includes the revised stories, we can train on the full dataset $\mathcal{D}^{tr} = D_1^{tr} \cup D_2^{tr}$, and see that the performance is uniform across the different test partitions. This highlights the challenges in constructing negative examples for such tasks and the importance of including contrastive examples for both training and test for proper generalization.

6.2 Story Revision for Counterfactual States

This generation task requires the model to revise a given story, such that the revised story is consistent with the given counterfactual participant state. We use human judgments to evaluate the revised stories because reference-based automatic evaluation metrics (BLEU [Papineni et al., 2002], BERTscore [Zhang et al., 2019] etc.) are inadequate for multiple reasons: (i) valid revised stories often exist that are different from the references, (ii) original and revised stories overlap heavily which can skew the metrics, and (iii) small lexical changes that don't change automatic metrics can affect logical consistency. We thus evaluate generation quality using our proficient workers from Section 3.2.

We compare performance of the models on a subset of 200 test instances chosen at random. We evaluated them for quality on three metrics: (1) **Inferable**: How likely is it for the given state α' to be true at any point in the revised story S'? This was rated on a 5-point Likert scale, which we thresholded to a 0/1 value (1 means inferrable). (2) **Logical**: Is the generated story S' logically correct? This was a YES/NO question. (3) **Minimal revison**: What is the degree of revision made to S to generate S'? This was rated on a 5-point Likert-scale, with 4 indicating minimal revision and 0 an entirely new story. Higher scores indicate higher similarity between S and S'. **Inferability** and **Logical** decide the ultimate correctness of a response. We calculated an overall model acceptability score (**ALL** in Table 6) by finding the percentage of model output that were both logical, and the input state can be inferred from them.

Table 6 shows that T5-l outperforms T5-b and GPT3 on the acceptability (**ALL**) of generated outputs by a large margin of 20% and 5.5%, respectively. GPT3 has the best performance on logical validity of the generated output with T5-l lagging behind by only 2%, but only 50% of GPT3's output satisfy the inferability criteria. In fact all the models have low inferability score, which brings down their overall acceptability score. T5-b has the best performance on the 'minimal revision' made to the original story, however this was not a primary metric of concern and there is always a trade-off between doing well on this score and generating an acceptable result. For example, revising a story conditioned on a counterfactual that is connected to entities in a different part of the story might require substantial revisions.

Overall, only 54% of the output generated by the best model, T5-l, are acceptable, indicating that the task is challenging and there is large room for improvement. Our results with GPT3 were based

Model	% Valid Attribute (A)	% Valid Inferable (B)	% Not in Story (C)	% ALL A, B & C
GPT3 FS	86.5	67.46	81.5	47.7
T5-b FT	96.75	41.0	90.25	35.17
T5-l FT	99.25	58.75	97.0	55.50

Table 7: State Change Generation - Human evaluation: % of output that satisfy the evaluation criteria. **Valid attributes** ensures that the generated states describe entity attributes, not actions. **Valid inferability** indicates if both states can be inferred from their correct stories. **Not in Story** assures that the states are unstated in the stories. **ALL** indicates the % of generations that satisfy all the these criteria.

	BERTscore	GLEU	rougeLsum
GPT3 FS	80.7	69.7	79.6
T5-b FT	81.6	73.2	81.7
T5-l FT	82.1	73.5	81.7

(a) Story Revision for Counterfactual States

	BERTscore	GLEU	ROUGEL
GPT3 FS	55.4	11.6	28.9
T5-b FT	54.4	11.7	29.5
T5-l FT	56.9	13.4	32.4

(b) State Change Generation

Table 8: Automatic evaluation for generative tasks: Model performance on the generative tasks using BERTscore, GLEU, and ROUGE based metrics.

on few-shot prompting where we treated its design choices as a modelling hyperparameter that were chosen based on automatic metric performance on the validation set. Few-shot performance of GPT3-scale models depends heavily on prompt engineering, so this direction may require further investigation.

6.3 State Change Generation

In this task, for given stories S and S', the model generates the two states α and α'. As in the previous task, we do a human evaluation of a randomly selected set of 200 model generated outputs.

Model outputs were evaluated on the following metrics: (1) **Valid Attribute:** Do the generated states α, α' describe entity attributes? This was a `YES` or `NO` question. (2) **Valid Inferability:** Are generated states α and α' inferable from S and S', but not from S' and S, respectively? Workers rated α and α''s likelihood of being inferred independently, on a 5-point Likert scale, which was thresholded to a 0/1 value. For instance, if α is inferred from S, then $L_{S\alpha} = 1$ (otherwise 0). Based on these scores, the inferability change for α is computed (1 for valid, 0 for invalid) using $max(0, L_{S\alpha} - L_{S'\alpha})$. (3) **Not in Story:** Are α and α' unstated in both S and S'? This was a multiple-choice question with 4 choices, 3 corresponding to the state being present in either one or both the stories, and the 4th for neither of the stories. An output (α, α') gets full credit on a metric if both α and α' are correct for that metric, half if only one of them (α or α') is correct, and 0 otherwise. **ALL** indicates full credit on all three metrics.

Table 7 shows the results. T5-l in general outperforms both T5-b and GPT3 on all the metrics except the **Valid Inferability**, where GPT3 outperforms the other models by a large margin. Interestingly, GPT3 is the worst performing model on **Valid Attribute** and **Not in Story**. This indicates that GPT3 is loosely ''cheating'' by copying text in the story itself, which of course is inferable, but violates the task's requirement of an implicit state. Overall, the best acceptability score (**ALL** in Table 7) is only 55.5%, which suggests that generating an output that satisfies all the criteria for a quality state change is an interesting challenge.

6.4 Automatic Evaluation for Generative Tasks

For the two generative tasks, we reported human evaluation results for the best analysis (prior sections). However, since human evaluation is expensive, we include here the results from three automatic metrics: GLEU (Wu et al., 2016), ROUGE (Lin, 2004), and BERTscore (Zhang et al., 2019). For GLEU, we consider 1 to 4-grams overlap between the output and reference. We report ROUGELsum for the Story Revision for Counterfactual States task since it is computed over the entire story, and the sentence level ROUGEL metric for State Change Generation. From Tables 8a and 8b, we can observe that even for automatic metrics, T5-l is still the best performing model on both tasks.

To further analyze automatic metrics as an alternative to human evaluation, we computed

	BERTscore	GLEU	ROUGE
Task-1	.21 (6e-7)	.14 (9e-4)	.13 (2e-4)
Task-2	.27 (4e-22)	.21 (1e-13)	.22 (1e-15)

Table 9: Pearson Correlation between automatic metric score of a model output and its validity as determined by humans. Numbers in parenthesis are p-value[6] for the null hypothesis that they are uncorrelated. **Task-1** is Story Revision for Counterfactual States, and **Task-2** is State Change Generation.

the correlation between them. We computed the Pearson correlation between the automatic metric score of an output and its validity as determined by humans. The results are reported in Table 9. The numbers in parenthesis are the p-values for the null hypothesis (95% confidence interval) that they are uncorrelated. We observed that BERTscore has the highest correlation with human evaluated validity for both tasks, outperforming other metrics by a substantial margin. The low p-value further indicates that the correlation is statistically significant. However, since the correlation is low, we strongly recommend using human evaluations, and only use BERTscore as an alternative where human evaluation is expensive.

6.5 Inter-Annotator Agreement

We measure the inter-annotator agreement (IAA) for the human workers using Gwet's Agreement Coefficient (Gwet, 2008, 2014), which is a type of generalized Kappa statistic.[7] Its interpretation is similar to generalized kappa (Viswanathan and Berkman, 2012), with $0.6 - 0.8 \equiv$ substantial and $\geq 0.8 \equiv$ almost perfect agreement. We use Gwet's coefficient because it is robust to the paradoxical behaviors (Wongpakaran et al., 2013; Gwet, 2014) seen in the commonly used IAA Kappa metrics (e.g., Cohen's and Fleiss). This paradoxical behavior of these metrics can lead to their IAA coefficients being lower even when the agreement is strong (Feinstein and Cicchetti, 1990; Byrt et al., 1993).

Crowd Workers IAA Table 10 shows the IAA coefficient for the tasks and their standard errors.

[6]The lower the p-value, the higher is the confidence for rejecting the null hypothesis.

[7]Gwet's normalizes the probability of observed agreement with a percent chance agreement that is the propensity of raters to agree on hard-to-rate instances (Gwet, 2014).

Task	Gwet's coefficient	
	Coeff	**StdErr**
Story State Inference	0.81	0.01
Story Revision from a Counterfactual	0.72	0.02
State Change Generation	0.76	0.01

Table 10: Inter-Annotator Agreement for human evaluation for the three tasks. **Coeff** is the calculated IAA coefficients, and **StdErr** is the standard error.

For each task, we computed the IAA coefficient for their respective evaluation metrics on their original scale (pre-thresholding[8]), which were then averaged to obtain the overall task scores. We computed the unweighted IAA coefficient for an evaluation metric if it was nominal, with quadratic weight if it was ordinal. As can be observed from the table, the crowd worker have strong agreement for both the generative tasks and almost perfect agreement for the classification task.

Experts IAA The two experts in Section 3.2 were responsible for accepting or rejecting a worker response for the PASTA creation. To measure their IAA, we created a pool of 200 PASTA instances that included both accepted and rejected instances. The experts had a Gwet's coefficient of 0.87 and agreed on 93.5% of those 200 instances.

7 Discussion

Here we discuss the main challenges and error analyses that highlight areas for future work.

7.1 Challenges

The key challenge common across all tasks is access to diverse types of knowledge (commonsense, numerical, factual, etc.), as well as the ability to combine and reason with them. For example, task 1 in Figure 3 requires factual knowledge about the temperatures at the North Pole, commonsense about snow and Christmas, and the ability to combine these when reasoning to detect the incompatibility of the input state.

The Story Revision Task has the added challenge of a model identifying the parts of the input story that are inconsistent with the counterfactual

[8]Note that thresholding was only done for ordinal scale metrics.

state, and then finally generating logically coherent text. For instance, in Figure 3 task 2, based on the input story and state, the model must first infer from sentences 2-4 that Connor had 12 coworkers. Then to generate the revised story, it also needs to reason about how the world state gets affected if there were fewer people than the number of doughnuts (e.g., now Connor would have some doughnuts left over).

The main challenge in the State Change Generation task is that there can be numerous plausible state pairs that are compatible with both stories, but they don't reflect a pertinent state change. Each state needs to be incompatible with one of the stories and compatible with the other, and this differentiation is a big challenge for any model. For example, in Figure 3 task 3, the observable difference between the stories is the outcome from coffee spilling on Joe. Using abductive reasoning with commonsense knowledge about temperature, one can easily infer that the change in state leading to a different ending comes from the coffee's temperature.

7.2 Error Analysis

We analyze the model's errors on 200 randomly selected instances from the validation set.

Story State Inference: We analyze model performance on different types of entity states following the categorization from Bhagavatula et al. (2019). We expand their spatial category to a broader set of physical attributes of entities (weight, temperature, location, etc.), and include a new Societal category to capture social constructs and norms. Even though multiple categories may apply to a state, to simplify our analysis we only use the most relevant category for each state.

In particular, we categorize each instance into one of the following: (i) Societal: knowledge about societal constructs such as relationship (*Jake is not married, I have 5 brothers*), norms (*John is not socially aware*), etc. (ii) Emotional/Psychological: knowledge about emotions (*John felt embarrassed, John hated Jake*), beliefs (*Jake believed in ghosts*), etc. (iii) Physical: Knowledge about physical attributes of entities (*Jake was in his school, the rock was very heavy, the coffee was hot, etc.*). Table 11 breaks down the overall performance of models across different categories. Models significantly under-perform on the societal category compared to the other two. In addition to the difficulty of modeling societal knowledge, we find that relatively more number of instances in this category require numerical commonsense, which adds additional complexity for the models. Physical commonsense is a broad category and its instances thus tend to cover a broad range of physical knowledge which could contribute to the difficulty of these instances. Emotional category has the best model performance since the inferred state include strong lexical indicators of emotions and feelings, similar to the observations in Bhagavatula et al. (2019).

	State Type	Acc. %	Contrastive Acc. %
	All - 100%	79	71.5
BERT-l	Societal - 14.5%	70.7	62.1
	Emotional - 54%	80.8	74.1
	Physical - 31.5%	79.8	71.4
	All - 100%	85.7	80.5
T5-l	Societal - 14.5%	81.9	75.9
	Emotional - 54%	87	81.5
	Physical - 31.5%	85.3	81
	All - 100%	90.6	86.6
RoBERTa-l	Societal - 14.5%	83.6	77.6
	Emotional - 54%	93.5	89.4
	Physical - 31.5%	88.9	86.5

Table 11: Story State Inference: Model performance for predicting the state inferability of different type of states.

The proposed generative tasks can have multiple correct outputs, each using a different set of commonsense knowledge. This makes it difficult to associate a unique knowledge category for the task instance. Therefore we manually analyze the outputs of the best performing model (T5-large) and identified common types of generation errors made by the model on the task.

Story Revision for Counterfactual States: The model output is correct for $\sim 58\%$ cases and incorrect for $\sim 42\%$. On analyzing the incorrect output, we found four main categories of error that we list in Table 12. The ''illogical revised story'' occurs when models produced revised stories that are logically incoherent (30% of errors). Generating logically coherent long text is still a challenging task for models, and to a certain extent can be attributed to their tendency to

Error Category	Percentage
Illogical revised story	30.1
Irrelevant change	27.7
Contradiction	20.5
Input state not entailed	20.5
State explicit in the revision	1.2

Table 12: Categories of different type of errors for the **Story Revision for Counterfactual States** task.

1. Irrelevant change
Input S: Al and Tara can't agree what to watch on TV. Al wants to watch soaps. Tara wants to watch the news. They play Paper, Rock, Scissors to decide. Tara wins and gets to watch what she wants. α': Al and Tara have two T.V.s
Output Al and Tara can't agree what to watch on TV. Al wants to watch soaps. Tara wants to watch the tv. They play Paper, Rock, Scissors to decide. Tara wins and gets to watch what she wants.
2. Illogical revised story
Input S: Rich was a musician. He made a few hit songs. Rich had a lot of fans who cared about him. He developed cancer. Eventually the cancer took his life and rich was gone. α': Rich was unknown.
Output Rich was a musician. He made a few hit songs. Rich had no fans. He became ill and lost his life. Eventually Rich recovered and began his recovery.
3. Contradiction
Input S: There was a large car wreck about two miles from the office. I was waiting for a co worker to relieve me. However, he was very late and I had no idea why. He called me a few minutes later. Apparently he was stuck in traffic because of the wreck. α': His co worker did not have access to a phone.
Output There was a large car wreck about two miles from the office. I was waiting for a co worker to relieve me. However, he was very late and I had no idea why. He called me on the phone a few minutes later but I had no idea why. Apparently he was stuck in traffic because of the wreck.
4. Input state not entailed
Input S: Frank was very skilled at shaping sculptures. His artwork started creating quite a buzz. A plain looking man approached Frank with a job. He wanted him to design the face for an adult themed product. Frank was about to decline until he saw the amount of the check. α': The check was for a very small amount of money.
Output Frank was very skilled at shaping sculptures. His artwork started creating quite a buzz. A plain looking man approached Frank with a job. He wanted him to design the face for an adult themed product. Frank was about to decline but didn't have much money.

Figure 4: Representative examples of top failure categories for the **Story Revision for Counterfactual States** task. Parts of the story that are changed are highlighted in pink, and the contradictory elements are highlighted in yellow.

forget attributes of specific entities (Welleck et al., 2018), ignore previously inferred facts (Sinha et al., 2019) and background information, or contradict previous statements (Brown et al., 2020).

Error Category	Percentage
Contradiction	37.4
Irrelevant states	35.2
Illogical state change	13.2
States reversed	4.4
No change in state	4.4
State directly stated in the story	4.4
Actions instead of states	1.1

Table 13: Categories of different type of errors for the **State Change Generation task**.

Moreover, 20.5% of the revised stories are categorized as *contradiction* as they clearly contradict the input counterfactual state. This corroborates previous findings on the challenges in reasoning about contradictions and negations (Hossain et al., 2020). Models also struggle to keep the changes relevant to the task criteria of the input state, which should be inferable from the revised story but not directly mentioned in it. They sometimes make *Irrelevant changes* (27.7% of errors) where they revise parts of the story that are not affected by the input counterfactual state. Other times they make revision that are inconsistent with the input counterfactual state (*Input state not entailed*, 20.5%) or the input *State is explicit in the revision* (1.2%), both of which do not meet the primary task requirements.

Figure 4 shows examples of the biggest error categories for the task.

State Change Generation: The model is correct for 54.5% of cases and fails for 45.5% when generating state changes. Table 13 shows the main error categories. While the model learns to generate both the α and α' states about the same entity, it makes many types of logical errors. *Contradictions* (37.4% of errors) are when a generated state is contradicted by its story, either directly or by deduction. *Illogical State Changes* (13.2%) are those where the generated states and input stories were topically related, but the states were simply illogical nonsensical. Both types of errors can be attributed to the challenges associated with making the relevant state inference, generating logically coherent text and reasoning about contradictions and negations. *Irrelevant States* (35.2%) are those where at least one of the generated states has no connection to its story. The error categories of *State Reversed* (4.4%), *No change in state* inferability (4.4%), *State is directly stated in the story*

1. Irrelevant states
Input S: Pat was looking for an interesting place to vacation. He searched various catalogs. Eventually he found the perfect destination. There he stayed in a small hut. It was one of the greatest experiences of his life. S': Pat was looking for a relaxing place to vacation. He searched various catalogs. Eventually he found the perfect destination. There he stayed in a luxury resort. It was one of the greatest experiences of his life.
Output α: Pat likes to be independent. α': Pat likes to be in a big city.
2. Illogical state change
Input S: Tom spent a long time writing an essay. He saved it on his memory card. When Tom went to print it the essay wasn't there. Tom had to go back to his computer. Luckily he was able to use a backup copy he made. S': Tom spent a long time writing an essay. He saved it on his memory card. When Tom went to print it the essay wasn't there. Tom had to go back to his computer. Unfortunately, Tom forgot to make a backup copy of the essay.
Output α: Tom is a good memory writer. α': Tom is a bad memory writer.
3. Contradiction
Input S: Rita accidentally left her smartphone on the plane. She felt terrible about this and knew her parents would be mad. Nobody ever turned in the phone, leaving her disappointed with people. Rita had to work many hours to buy a new phone as a replacement. She is significantly more careful with her belongings now. S': Rita accidentally left her smartphone on the plane. She felt terrible about this and knew her parents would be mad. Nobody ever turned in the phone, leaving her disappointed with people. Rita quickly bought a new phone as a replacement. She is significantly more careful with her belongings now.
Output α: Rita was unemployed α': Rita was rich

Figure 5: Examples of top failure categories for the **State Change Generation** task. For the contradiction category, the conflicting parts are highlighted.

(4.4%), and outputs are *Actions instead of states* (1.1%) are due to models' inability to correctly understand the task constraints. Figure 5 shows example of the major error categories for the task.

7.3 Interactive Feedback with LLMs

Based on the error analysis for the tasks performed above, the majority of the error categories can be attributed to the model's inability to maintain factual and logical consistency in the generated output. For the Story State Inference task, the lack of consistency is further demonstrated by the low contrastive accuracy on the task. Conversation-based LLMs such as ChatGPT (OpenAI, 2022) or LaMDA (Thoppilan et al., 2022), have been shown to have both knowledge at the scale of LLMs such as GPT3 and an ability to incorporate human feedback for NLU tasks. These capabilities may enable them to leverage feedback about inconsistencies (if detected) in the initially generated output to correct these inconsistencies in the subsequent generations. However, when the task is to be performed at scale, the feedback that guides the model to the correct output needs to be automatically generated instead of a human guiding the model. As such, this type of model presents a fruitful and challenging research direction to address some of the issues and further improve performance on the tasks.

8 Conclusion

In this work, we introduced a new resource, PASTA, that captures unstated commonsense knowledge required to understand and reason about participant states in a narrative. PASTA opens the door to developing more complex reasoning abilities, especially those that require access to implicit information. We described three PASTA reasoning tasks, one classification and two generation, that test for different aspects of state-based reasoning. This work shows that with careful crowdsourcing and contrastive design we can obtain a high-quality dataset that can be used to evaluate deeper reasoners. Benchmarking results suggest that PASTA tasks are not within the reach of current large sized models, as of yet, and encourages future research in modeling commonsense knowledge with states.

Acknowledgments

We would like to thank the anonymous reviewers for their comments, questions, and suggestions. This material is also based on research that is in part supported by the NSF, Grant No. 2007290, Army Research Laboratory, Grant No. W911NF2120076, and by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright 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 the Air Force Research Laboratory (AFRL), DARPA, or the U.S. Government. This material is based in part upon work supported by the National Science Foundation under grant no. IIS-2024878.

References

Niranjan Balasubramanian, Stephen Soderland, Mausam, Oren Etzioni, et al. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. Simulating action dynamics with neural process networks. *ArXiv*, abs/1711.05313.

Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. 2020. Cue me in: Content-inducing approaches to interactive story generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 588–597, Suzhou, China. Association for Computational Linguistics.

Petter Bae Brandtzaeg, Asbjørn Følstad, and Maria Ángeles Chaparro Domínguez. 2018. How journalists and social media users perceive online fact-checking and verification services. *Journalism Practice*, 12(9):1109–1129. https://doi.org/10.1080/17512786.2017.1363657

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429. https://doi.org/10.1016/0895-4356(93)90018-V, PubMed: 8501467

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610. https://doi.org/10.3115/1690219.1690231

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. In *NAACL*. https://doi.org/10.18653/v1/N18-1144

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549. https://doi.org/10.1016/0895-4356(90)90158-L, PubMed: 2348207

Francis Ferraro and Benjamin Van Durme. 2016. A unified bayesian model of scripts, frames and language. In *Proceedings of the 30th Conference on Artificial Intelligence (AAAI)*, pages 2601–2607, Phoenix, Arizona. Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v30i1.10328

Álvaro Figueira and Luciana Oliveira. 2017. The current state of fake news: Challenges and opportunities. *Procedia Computer Science*,

121:817–825. https://doi.org/10.1016/j.procs.2017.11.106

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Antony Galton. 1990. A critical examination of allen's theory of action and time. *Artificial Intelligence*, 42(2–3):159–188. https://doi.org/10.1016/0004-3702(90)90053-3

Zahra Ghadiri, Milad Ranjbar, Fakhteh Ghanbarnejad, and Sadegh Raeisi. 2022. Automated fake news detection using cross-checking with reliable sources. *arXiv preprint arXiv: 2201.00083*.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: An interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 89–97, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4016

Kilem L. Gwet. 2014. *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48. https://doi.org/10.1348/000711006X126600, PubMed: 18482474

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5516

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia. Association for Computational Linguistics. https://doi.org/10.3115/1220175.1220289

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.732

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*. https://doi.org/10.1609/aaai.v35i7.16792

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.deelio-1.10

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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. https://doi.org/10.18653/v1/N16-1098

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.370

OpenAI. 2022. Chatgpt. https://openai.com/blog/chatgpt

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. https://doi.org/10.3115/1073083.1073135

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *EMNLP*. https://doi.org/10.18653/v1/D19-1509

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1213

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035. https://doi.org/10.1609/aaai.v33i01.33013027

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157.

Lei Sha, Sujian Li, Baobao Chang, and Zhifang Sui. 2016. Joint learning templates and slots for event schema induction. *arXiv preprint arXiv:1603.01333*. https://doi.org/10.18653/v1/N16-1049

Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Bicharra Garcia. 2019. Can machines learn to detect fake news? A survey focused on social media. In *Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2019.332

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1458

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v31i1.11164

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for ‘‘what if. . . ’’ reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1629

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark,

Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.520

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2948–2958, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1302

Meera Viswanathan and Nancy D. Berkman. 2012. Development of the rti item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology*, 65(2):163–178. https://doi.org/10.1016/j.jclinepi.2011.05.008, PubMed: 21959223

Sean Welleck, Jason Weston, Arthur D. Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. In *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1363

Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: From general language models to commonsense models. In *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2022.naacl-main.341

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(1):1–7. https://doi.org/10.1186/1471-2288-13-61, PubMed: 23627889

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1010

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.