POQue: Asking Participant-specific Outcome Questions for a Deeper Understanding of Complex Events

Sai Vallurupalli[1], **Sayontan Ghosh**[2],
Katrin Erk[3], **Niranjan Balasubramanian**[2], **Francis Ferraro**[1]
[1] University of Maryland, Baltimore County,
[2] Stony Brook University,
[3] University of Texas, Austin
kolli@umbc.edu, sagghosh@cs.stonybrook.edu,
katrin.erk@utexas.edu, niranjan@cs.stonybrook.edu, ferraro@umbc.edu

Abstract

Knowledge about outcomes is critical for complex event understanding but is hard to acquire. We show that by pre-identifying a participant in a complex event, crowdworkers are able to (1) infer the collective impact of salient events that make up the situation, (2) annotate the volitional engagement of participants in causing the situation, and (3) ground the outcome of the situation in state changes of the participants. By creating a multi-step interface and a careful quality control strategy, we collect a high quality annotated dataset of 8K short newswire narratives and ROCStories with high inter-annotator agreement (0.74-0.96 weighted Fleiss Kappa). Our dataset, POQue (**P**articipant **O**utcome **Que**stions), enables the exploration and development of models that address multiple aspects of semantic understanding. Experimentally, we show that current language models lag behind human performance in subtle ways through our task formulations that target abstract and specific comprehension of a complex event, its outcome, and a participant's influence over the event culmination.

1 Introduction

Situations that people experience or describe can be complex, and developing a computational understanding of these situations is not straightforward. Consider the short narrative from Fig. 1:

> After a decade as renters, [the Brofmans] were finally able to buy a small house here four years ago. But if the Argentine government yields to [IMF] pressure to rescind emergency legislation meant to protect ordinary families like the Brofmans, the couple stand to lose their home and the $32,000 they have paid for it so far.

Across multiple, interwoven events with multiple participants, this narrative describes part of the process of losing one's house. A *possible* ending (the loss of a home) is suggested, which is the result of a confluence of these events. This ending can be semantically grounded in various changes of state that the participants experience, though note how the use of counterfactual considerations, conditional statements ("if the Argentine government..."), and varying levels of certainty over whether events have actually happened (e.g., realis vs. irrealis) contribute to the difficulty in understanding this complex event (Herman, 2002; Ryan, 1991).

Knowledge about how event outcomes affect individual participants can help identify salient events in a narrative, fill in implicit missing information (LoBue and Yates, 2011) and chain events that lead to improved understanding of complex events (Graesser et al., 1994). To infuse AI models with similar knowledge, narrative comprehension research has focused on learning event relationships (Mostafazadeh et al., 2016; Chambers and Jurafsky, 2008; O'Gorman et al., 2016; Caselli and Vossen, 2016), using temporal (Pustejovsky et al., 2003), causal (Mirza et al., 2014) and discourse (Prasad et al., 2008) relationships in text. However, as Dunietz et al. (2020) and Piper et al. (2021) argue, for a more useful, generalizable, and robust comprehension, we need to take a holistic view of complex events. In this paper, we tackle an understudied notion of this holistic view and examine knowledge of post-conditions based in states to support inferences of the form "who did what to whom and with what end result."

A core insight we make is that viewing complex events through the lens of a single participant at a time, either from an agent (how the participant affects others) or a patient (how the participant is affected) view, can help mitigate the complexities we have discussed. In this we build on cross-disciplinary research that shows that humans mentally structure events along single participants (Black and Bower, 1980; Morrow et al., 1989), and that participant-based event and outcome analysis improves complex event understanding (Dijk and Kintsch, 1983; Liveley, 2019). We

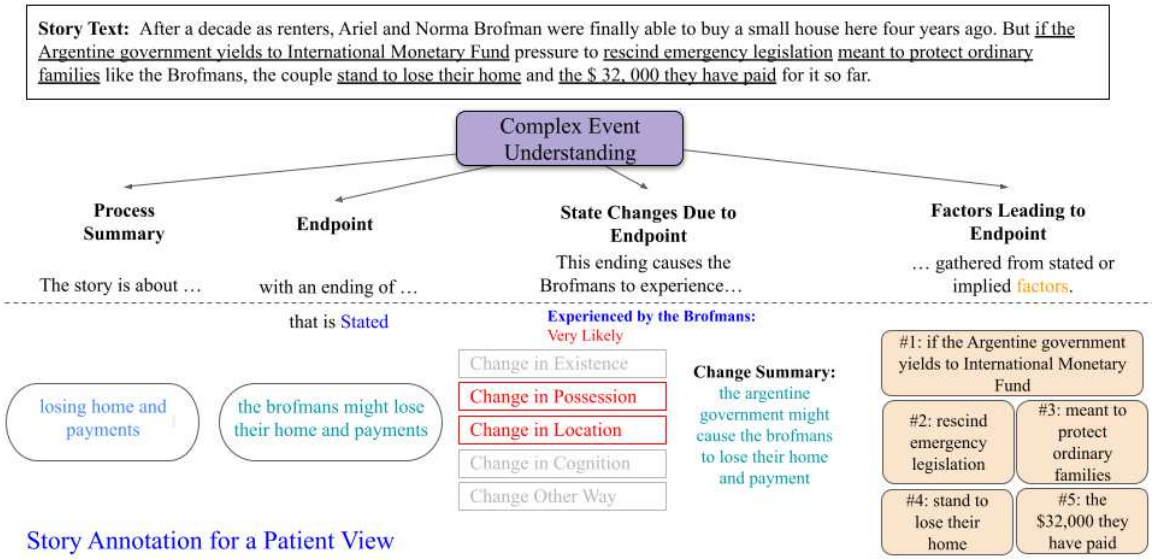

Figure 1: Our approach to understanding state change outcome of complex events. Our four step annotation process involves describing abstractly what a story is about; writing an endpoint for the story; identifying and describing the changes that are a result of the endpoint, and identifying the salient sub-events that lead or support to those changes. To mitigate this complexity, we focus annotator's attention on one particular participant, and how that participant either causes or experiences the identified changes. This process has resulted in 4k annotated documents.

also note that a complex event is not exhaustively described by what is stated in the text: it is well known that speakers often omit narrative steps that can be inferred (Grice, 1975), including outcomes and effects of a narrative that are often left implicit.

We present POQue, a dataset with post-conditional knowledge about complex events. We identify cumulative outcome-oriented endpoints of the stories caused by related events and the consequences or post-conditions of those events as state-based changes in participants. Seen in Fig. 1, in a storyline involving a participant ("the Brofmans"), we identify an ending outcome for the complex event (the *Endpoint*, "the brofmans might lose their home and payments"), with salient events that lead to this (*Factors Leading to the Endpoint*). We relate a participant's involvement in the complex event (as a "patient" who "Very Likely" experienced the ending outcome) and the changes of state occurring as a result of the complex event (the changes in possession and location experienced by the Brofmans and other families, and the change in possession by the Argentine government and IMF).

To facilitate high quality annotations we designed a multi-stage crowd sourcing solution to acquire, monitor, assess and curate annotations at scale. We collected 7772 annotations across 4001 stories and assessed a random 1545 annotations (20%) in a multistage pipeline to obtain a highly curated test set. Using POQue, we test current language models on reasoning about complex events in narratives: we formulate challenge tasks to identify and generate post-conditions from a story, and evaluate how well trained models predict a participant's involvement in enabling a complex event.

We summarize our contributions as follows: (1) we introduce a new annotation scheme focusing on complex events from the point of view of a single participant. (2) We create a new dataset of complex events from three collections of everyday stories, using free form text to obtain insight into implied outcomes. (3) We obtain high quality annotations from crowd workers without the use of requester generated qualification tests. (4) We formulate challenge tasks aimed at evaluating the ability of language models to perform richer complex event comprehension, specifically: a) generating a process summary of the complex event b) generating an endpoint of the complex event, c) generating the outcome of a complex event based on a participant's semantic role d) identifying a participant's involvement in a complex event, and e) generating post-conditions or changes caused by a complex event. Our dataset and code are publicly available at https://github.com/saiumbc/POQue.

2 Related Work

Narrative texts communicate experiences and situations by connecting related events (Brooks, 1984; Mateas and Sengers, 1999) through events involv-

ing participants (Bal, 1997; Eisenberg and Finlayson, 2017; Liveley, 2019). Previous works, viewing narratives as sequences of events, annotated event pairs for event coreference, temporal, and causal relationships (O'Gorman et al., 2016; Caselli and Vossen, 2016; Mirza and Tonelli, 2016; Mostafazadeh et al., 2016). Newer works have studied event groups using predicate hierarchies (Qi et al., 2022) and temporal graphs (Li et al., 2021). However, these approaches focus on event-event relationships, without diving deeply into participant or entity analysis. Unsupervised methods assume narratives are coherent and learn partially ordered event chains (Chambers and Jurafsky, 2008; Balasubramanian et al., 2013) or sub event relationships (Yao et al., 2020) but these are limited to what occurs in the text itself, which can lead to well-known issues of bias or evaluation limitations (Gordon and Van Durme, 2013; Rudinger et al., 2015).

Caselli and Inel (2018) obtain crowd annotations of causal relationships between events and assess their quality by relating them to expert annotations. PeKo (Kwon et al., 2020) uses crowd annotations of precondition relationships and fine tunes a language model for finding such relationships. However, both works limit their study to event pairs in short text snippets. The ESTER dataset (Han et al., 2021) consists of more comprehensive relationships in a story, though limited to within-text (i.e., stated) mentions of events only. GLUCOSE (Mostafazadeh et al., 2020) provides elaborate causal relationships for several event dimensions for each event in a story. However, in contrast to our effort, these works address direct causal relationships between within-text events and do not focus on participants.

Understanding complex events has long attracted cross-disciplinary attention. For example, theoretical linguistics and cognitive science work has shown that humans understand a narrative text using simulative inference (Kaplan and Schubert, 2000; Boella et al., 1999; Schubert and Hwang, 2000). Prior work has also shown how observing participants' events and the resulting consequences can lead to improved understanding of events (Dijk and Kintsch, 1983; Zwaan and Radvansky, 1998).

3 Knowledge Representation

As an underlying motivation for our efforts, we posit that for language models to be able to reason about complex events from narratives, they should be able to identify a likely ending of that complex event, component events that lead to that ending, and the state changes that result from that ending. However, this type of knowledge is complex and has been computationally understudied, leading to a scarcity of sizable datasets. In our efforts to correct this, we appeal to classic, cognitively- and linguistically-backed results.

First, inspired by the idea from Kintsch and van Dijk (1978) that text comprehension involves reducing relevant details into an abstract coherent semantic text, our targeted annotations include an abstract high level summary and the minimal set of salient events that make up the complex event. Second, we extend the idea of thematic roles for verbal arguments (Dowty, 1991) to a generalized semantic role for the complex event. Specifically, Dowty showed that easily verifiable characteristics and properties, such as volitional participation in an event or whether a participant underwent a change of state because of a *particular* event, can be used to define predicate-level prototypical semantic roles. Inspired by this, we characterize the roles of complex events through the intentional engagement in changes of state of participants. As such, our targeted annotations account for both the intentional involvement of the participants in the complex event and the cumulative impact of all the events that make up the complex event.

Story and Participant Stories in our dataset are either a ROCStory or heuristically salient portions of newswire (first 100-150 tokens).[1] For more information on story processing see §4.1. We define a participant as an entity that was mentioned several times in the story. See §4.1 for more on participant selection. Multiple entities are considered a single Participant if these entities are mentioned together and participate together in all the events. In Fig. 1, "the Brofmans" are a Participant.

3.1 Targeted Knowledge Annotations

Given a story S, a participant P, and P taking on a agent-like or patient-like cumulative semantic role, PR, we obtain the following annotations.

Process Summary (PS): A high-level, free-form description of the situation, which provides the topical context for the complex event. For example, "Losing home and payments" for the story in Fig. 1.

[1] While we acknowledge they have important differences, we use "narrative" and "story" interchangeably.

Endpoint Description (ED): A free-form description of the inference of what happened or is likely to happen in the story, conditioned on the process summary. It is the result of an aggregate of story events that leads to state changes for participants. For the story in Fig. 1, the endpoint is "The Brofmans might lose their home and payments."

Endpoint Anchoring (EA): An endpoint may be (inferentially) *stated* in the text, or it may be *implied*/suggested. We judge this via a three-way choice (stated, implied, unsure). For the story in Fig. 1, the endpoint is *stated* by the text.

Participant Involvement (PI): Whether the participant caused the endpoint, or experienced it, indicated with a 5 point Likert scale, from very unlikely to have caused [experienced] the endpoint, to very likely to have caused [experienced] it. For the story in Fig. 1, the complex event maximally affects P. Hence this rating is a "very likely."

Change Summary (CS): A templated text description of state changes caused or experienced by participants as a result of the endpoint. In Fig. 1, this is "The Argentine government might cause the Brofmans to lose their home and payment."

Change Modes ($(\mathbf{c_1}, \mathbf{c_2}, .., \mathbf{c_k})$): The various ways in which participants experience changes. These change modes are: change in existence, feeling, location, possession, some other way, or no changes. This list was inspired by classic linguistics, c.f., Dowty (1991), though refined during early examination of our stories.

Factors ($(\mathbf{f_1}, \mathbf{f_2}, .., \mathbf{f_n})$): Salient events that lead to the endpoint and state changes where each factor captures an event in a phrase comprising of at least a subject and verb. For the story in Fig. 1, "'if the Argentine government yields to International Monetary Fund," is one of the factors leading to the Endpoint.

3.2 Crowd Annotations

We created a human intelligence task (HIT), deployed on Amazon Mechanical Turk (AMT). The HIT consists of a story with highlighted participant mentions displayed in the left column and four annotation steps in the right column which vary slightly for the agent and patient views. The protocol was IRB approved.

Crowd workers were instructed to read the story, focus on the highlighted participant, and provide annotations. We provide several annotated examples, general instructions for completing the HIT, and specific instructions for each step suggesting a template to follow for some steps. More details and the layout of the HIT are in Appendix B.

The annotation task consists of 4 steps and each story is assigned to two workers, one where the highlighted participant is assigned the role of "agent" and another with the assigned role as "patient." Step 1 asks for a high level description of the story, a process summary of the situation described. Step 2 of the HIT asks for a description of an endpoint in the story. We assume a story's endpoint typically signifies a state change caused by a complex event. Step 3 asks for a summary of changes caused by the complex event in the story participants and also asks to identify the type of changes. Step 4 of the HIT asks an annotator to identify the salient events, or factors, that lead up to the complex event and changes from it.

4 Dataset

We selected stories from three narrative English language datasets – the ESTER dataset (Han et al., 2021), the ROC stories dataset (Mostafazadeh et al., 2016) and passages from the Annotated New York Times newswire dataset (Sandhaus, 2008). We selected these given the prominence the underlying documents have in the broader NLP community (the ESTER documents are a subset of the TempEval3 (TE3) workshop dataset (UzZaman et al., 2013). We included ROC stories because they often contain a single situation with mostly salient information. We noticed these stories help crowd workers easily focus on salient events, providing cues for factors and state changes. Meanwhile, ESTER and the selected Newswire stories provide a variety of complex situations and discourse text. Additionally, by selecting subsets of these well-known datasets, we hope that future efforts may be able to aggregate our annotations with existing ones, enabling richer phenomena to be examined.

4.1 Story Preparation

We sampled stories from the Annotated New York Times (ANYT) corpus, ROCStories, and ESTER. We then identified participants via an automatic entity coreference system. We heuristically selected relevant and annotable excerpts of the document by identifying "continuant story lines" (see Appendix A). After identifying a participant and a

	# Stories	# Agent	# Patient
Total	4001	3896	3876
ROC	1383	1364	1356
ESTER	1275	1237	1218
NYT	1343	1295	1302

Table 1: Document-level data statistics. Note that the number of stories refers to the number of unique stories annotated, while the agent and patient numbers refer to the number of instances annotated on those documents. Additionally, 260 of the ROCStories are from the CATERS (Mostafazadeh et al., 2016) collection. CATERS stories and ESTER stories containing subevents are useful for relating causal and compositional events in a complex event.

Avg. process summary length	5.7 words
Avg. endpoint length	9.4 words
Endpoint stated/implied/unsure %	68.5%/28.9%/2.6%
Avg. change descr. length	8.9 words
Avg. likelihood of causing change (agent)	4.0 (likely)
Avg. likelihood of experiencing change (patient)	4.1 (likely)
Avg. # of factors	3.7
Avg. factor length	8.0 words

Table 2: Additional statistics about POQue.

continuant story line, we randomly selected 4001 stories for annotation; see Table 1 for details.

4.2 Dataset Annotation & Pricing

Our dataset has 4001 stories, annotated by 163 different crowd workers. The average number of stories annotated per worker is 43. When possible, we annotated from both an agent and a patient view for a participant, so in total we obtained 7773 annotations. Workers were paid an average of $0.50 for annotating a single HIT, either an agent or a patient view of the story. For more detailed information on payment and training see Appendix B.3. We tackle the positivity bias in AMT work (Matherly, 2018) using a thorough initial verification and training (see Appendix B) and ensured workers understood the task and provided quality annotations.

4.3 Dataset Statistics

For most stories we obtained two annotations, with the two participant semantic roles. We show high-level document statistics in Table 1. Annotations for the two different roles of the highlighted participant are shown in separate columns. We show more detailed annotation statistics in Table 2, and examine the frequency of change modes in Fig. 2.

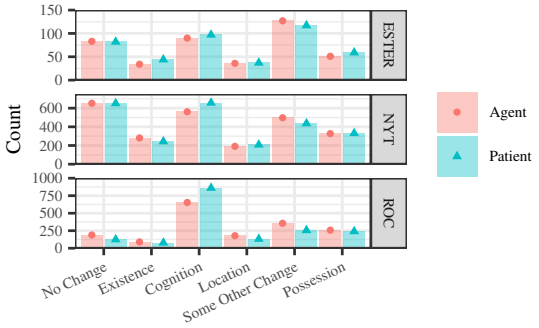

Figure 2: Count of change modes, shown for each of the agent and patient roles, and broken out across the originating datasets our annotated narratives come from.

4.4 Dataset Quality and Analysis

We noticed that the nature of the stories and the task steps elicit a variation in the text style and format, even from the same worker. Our experiments (§6) and ablation studies (§7) did not uncover any easy biases attributable to a small number of workers producing most of the annotations. Due to space limitations we explain our process for evaluating 1545 random annotations in Appendix C.2.

5 Tasks for State Change Knowledge

Based upon our collected dataset, we propose several tasks. These tasks are designed to test various aspects of comprehension involving complex events, their participants, and outcomes.

5.1 Task 1: Generating Process Summaries

Categorizing stories based on the type of situation they describe is necessary for generalization. For this, we fine-tune models to generate an abstract and high level process summary of the complex action described in a story. Because we annotated salient events for the story, i.e., the factors, we have two task formulations. We fine-tune models to generate PS either given S or $(f_1, f_2, .., f_n)$. Both are standard summarization tasks which we compare with a baseline where the process summary is assumed to be "About P."

5.2 Task 2: Generating Complex Event Endpoints and Salient Events

Understanding a story involves the identification and decomposition of salient events that lead to an endpoint, for the described complex event. We test this understanding with two complementary formulations where we generate either the endpoint description or the salient events, i.e., factors. For generating ED we have two sub formulations where

we fine-tune models either on S or $(f_1, f_2, .., f_n)$. For generating $(f_1, f_2, .., f_n)$ we fine-tune models on (S, ED). These are all standard summarization tasks which we compare with a baseline where ED is assumed to be the last sentence of the story.

5.3 Task 3: Generating Changes Resulting from a Complex Event

Knowing the changes caused by a complex event gives us an insight into its importance and the intentions (addressed in Task 5) behind it. In this task, we generate changes caused by a complex event through the lens of the semantic role tracking we have employed throughout our effort. Using standard summarization, we fine-tune models to generate CS given (S, ED, PR).

5.4 Task 4: Identifying Types of Changes

Grounding the impact of a complex event in the various change modes a participant undergoes helps in understanding the importance of new situations by relating them to known situations with similar post-conditions. We formulate this as a multi-label binary classification and fine-tune models to identify $k = 5$ change modes $(c_1, c_2, .., c_k)$ given S.

5.5 Task 5: Assessing Participant's Involvement in the Complex Event

Besides the story context, the participant's semantic role heavily influences our decision of whether the participant intended or enabled the complex event or the changes caused by it. In this binary classification task, we predict the participant involvement rating PI given (S, ED, PR). This prediction demonstrates a model's ability to identify a participant's intentional engagement and enablement of the complex event and its impact.

6 State Change Benchmark Experiments

We benchmark the performance of current encoder-decoder transformer language models, T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), which are effective for both text generation and classification. We compare fine-tuned base and large[2] models with multiple automated metrics and crowd sourced human evaluation. We use bootstrap for calculating statistical significance via the `mlxtend` library (Raschka, 2018).

[2]We found that training a bart-large model was finicky, with some training runs not converging. For these cases, we do not include results for the bart-large model. See https://github.com/huggingface/transformers/issues/15559.

6.1 Automated Evaluation

We use the classic metrics of **ROUGE** (Lin, 2004), **BLEU** (Papineni et al., 2002), and **METEOR** (Lavie and Agarwal, 2007), and the more recent **BertScore** (Zhang et al., 2020). Due to space limitations, we present ROUGE-L and BertScore in the main paper, and additional ROUGE-1, ROUGE-2, METEOR, and BLEU scores in the appendix (Appendix F). We use standard metrics used for single and multi-label classification: **Accuracy** and **macro F1**. In multi-label classification, we calcuate **Subset Accuracy** and **macro F1** using sklearn and a **Hamming Score** which is computed as $\frac{1}{n}\Sigma_{i=1}^{n}\frac{Y_i \cap Z_i}{Y_i \cup Z_i}$, where Y and Z are true and predicted labels for n examples.

6.2 Human Study of Model Generations

We perform a human evaluation of the generation tasks (1, 2, and 3) using 50 randomly selected generations for each model and the corresponding human annotations. We obtained qualitative ratings from 3 crowd workers experienced in annotating our HITs and measured IAA using a weighted Fleiss's Kappa as in Appendix C. For each summary, workers are presented with the story and the summary and asked to rate the summary on aspects that relate to the task such as abstractness, factuality and salience using a 5-point Likert scale. See Appendix D for more information on these aspects and the HITs used for evaluation.

6.3 Task 1: Generating Process Summaries

To test whether a model generates a more focused process summary when trained on salient information, we compare pre-trained models fine-tuned on S and $(f_1, f_2, .., f_n)$ with an easy baseline process summary of "About P," where P is the participant. Less than 1% of the process summaries in dataset and model generations contain this baseline format. Results from this task training are listed partially in Table 3 and fully in Table 11. For all models, Rouge, BLEU and METEOR scores show less lexical overlap, but BertScore indicates a high similarity between the model generated and reference summaries. Inspired by previous work in measuring abstractiveness (See et al., 2017; Dreyer et al., 2021; Gao et al., 2019; Narayan et al., 2018), we compare average number of tokens (*Len*) across all summaries, the percentage of exactly matched trigram spans in the story (*Ext*), and the average of Abstractness Likert scores (*Abstr.*) from the evalua-

	Model	Len	Ext. (↓)	RougeL	BertScr	Abstr.
	Reference	**3.6**	**.13***	-		**3.57***
	About P	1.7	.27	10.43	83.86	2.37
Story	Bart-base	4.0	.46	21.43	86.70	2.77
	T5-base	10.0	.60	19.50	85.99	2.32
	T5-large	6.9	.63	20.30	86.15	2.13
Fact.	Bart-base	4.2	.33	**23.81**	**87.66**	3.22
	T5-base	9.9	.56	18.29	86.10	2.73

Table 3: Generating Process Summaries (Task 1). See appendix Table 11 for the full results. Bart-large is not included because we were unable to get it to properly converge. The best scores are bolded. We use * to indicate a significantly higher value than other values in the column with a p value between 0.001 and 0.0001 (except for Bart-base trained on Factors where the p value is 0.13). *Len* value for Reference is considered the best as the baseline value is not meaningful.

tion HIT (see Fig. 17) for process summaries. We provide Len, Ext and Abstr. values to help contextualize scores. While a lower value of Ext does not necessarily mean a better generation, it does mean there is less direct copying of length-3 phrases.

Discussion: BART generations are brief, less extractive and more abstractive, whereas T5 generations are longer, less abstractive and more directly drawing upon spans of story text. The Reference summaries are brief, significantly less extractive and at a significantly higher abstractness score compared to all the models. Models fine-tuned only on the factors produce more abstractive summaries. However, this increase in the abstractness for the BART model increased the factual errors, in line with previous observations (Cao et al., 2017; Kryscinski et al., 2019; Dreyer et al., 2021). The significantly higher brevity and abstractness of the Reference summaries point to a substantial gap between human and LMs' ability at capturing complex actions in a brief, high-level, abstract phrase.

6.4 Task 2: Generating Endpoints & Factors

To test how well models generate endpoints, models are fine-tuned to generate ED, given S or $(f_1, f_2, .., f_n)$ and compared with the baseline version where the ED is assumed to be the story's last sentence. Partial results for this task formulation are listed in Table 4 and the full results in Table 12 for the trained models. We also compare models trained on the complementary task of generating $(f_1, f_2, .., f_n)$ given (S, ED). A special token separates factors in all the task formulations involving factors. Partial results for this complementary task are listed in Table 5 with the full results in Table 13.

For all models, Rouge, BLEU and METEOR scores show higher lexical overlap, and BertScore

	Model	Len	Ext↓	RougeL	BertScr	Fact.	Sal.
	Reference	7.9	.27	-		4.15	3.46
	Last sent.	23.3	.79	21.82	85.60	4.49	3.35
Story	Bart-base	11	.72	25.43	87.61	4.66	3.97
	Bart-large	10.3	.63	24.74	87.62	4.59	3.81
	T5-base	13.6	.70	24.07	87.19	**4.71***	4.03
	T5-large	12.9	.67	**25.71**	87.54	**4.71***	**4.23***
Fact.	Bart-base	7.3	.49	24.09	**87.93**	4.11	3.28
	T5-base	10.4	.47	22.01	87.07	3.99	3.02

Table 4: Generating Endpoints for stories and factors (Task 2a). See appendix Table 12 for the full results. The best scores are bolded. * indicates the value is significantly higher than the Reference value with a p value of .002 for Factuality and .0008 for Salience.

Model	# fact.	Len	RougeL	BertS	Brev.	Fact.	Sal.
Reference	3.6	8.3	-		**3.35***	3.25	3.04
Bart-base	3.5	14.0	45.28	88.06	2.54	3.49	3.57
Bart-large	3.6	13.7	45.98	88.10	2.12	3.2	3.31
T5-base	2.6	19.6	43.31	87.74	3.23	3.69	**4.01***
T5-large	3.7	13.9	**47.96**	**88.44**	2.85	**3.80***	3.96*

Table 5: Generating Factors from stories and their endpoints (Task 2b). See appendix Table 13 for the full results. The best scores are bolded. * indicates the value is significantly higher than the Reference value for Factuality and Salience with a p value of .0001. The Reference value for Brevity is significantly higher than all values in the column with a p value of .0001.

indicates a high similarity between the generated and reference summaries. We compare the average of the Factuality and Salience scores (*Fact.* and *Sal.*, resp.) from the endpoint summary evaluation HIT (see Fig. 18) along with the average number of tokens (*Len*) across all summaries and the percentage of exactly matched trigram spans in the story (*Ext*). We also compare the average of the Brevity, Factuality and Salience scores (*Brev.*, *Fact.*, and *Sal.*) from the evaluation HIT (see Fig. 20) for factors along with the average number of factors (*# fact.*) across all stories and the average number of tokens (*Len*) across all factors and stories.

Discussion: Scores are significantly higher for LM generations than Reference endpoint descriptions on both Factuality and Salience. We looked at a random 100 Reference endpoint descriptions and the corresponding model generations. The first author of this paper identified which of the endpoints were not directly stated in the story, but rather implied by the story, and which ones were not-factual. As shown in Fig. 3, very few of the model generations are implied endpoints and a third of the Reference endpoints contain implied descriptions. Our HIT instructions asked workers to annotate not only explicit endpoints but also the ones implied by the story and they identified 29% of them as implied. Evaluators lowered their scores both

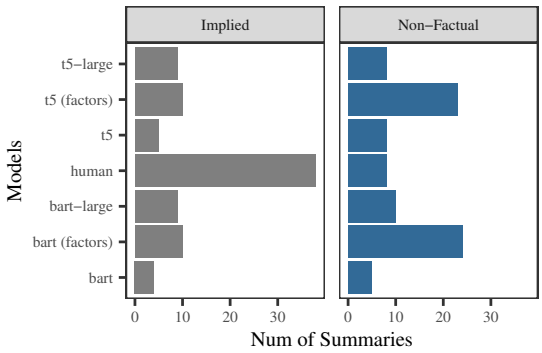

Figure 3: Non-factual and implied endpoint types.

for factuality and salience for these implied endpoints as the description may be a possible but not strictly entailed outcome. Models trained on factors generated more implied endpoint descriptions but these implied endpoints contained more factual errors possibly because of less available context. We conclude that LMs try to generate stated endpoint descriptions unless challenged by limited context.

Fine-tuning on stories and endpoints generates salient factors, indicated by the high assessment scores. However, the generated factors on average contain multiple facts making them less concise and focused than human written factors.

6.5 Task 3: Generating Change Summaries

We compare model generations of state changes, by fine-tuning models to generate CS conditioned on (ED, PR). Given the pair (S, t), where t is either "P caused this: ED" if PR="agent" or "P experienced this: ED" if PR="patient", fine-tuned models generate CS. Partial results for this task are listed in Table 6, and the full results in Table 14. For all models, Rouge, BLEU and METEOR scores show high lexical overlap and BertScore indicates a high similarity between model generations and Reference summaries. We compare the average of the Factuality and Salience Likert scores (*Fact.* and *Sal.* resp.) from the evaluation HIT (see Fig. 19) which measures whether the generated text contains change(s) resulting from the complex action.

Discussion: T5 models generate change summaries that are significantly more factual and salient than the BART models. While T5 generations score higher than Reference summaries, the difference is not significant. To explain these results, we inspected the 50 evaluated stories and found that less than 10% of the stories have no changes even though annotators indicated "no changes" for 25% of the stories in these 50 (and in the entire dataset). To see if crowd workers can identify these no-change stories, we ran a HIT where the change summary for these 50 stories was set to no-changes. From the results of this HIT we found that human evaluators also think there are 3 times as many stories with no-changes. Workers missed subtle changes in a story especially when they relate to changes in cognition. T5 was able to identify story text that contained subtle changes while the BART models seem to be learning the data distribution from the training data. BART generations also contain a higher number of factual errors leading to its subpar performance.

Model	RougeL	BertScr	Fact.	Sal.
Reference	-		3.36	3.32
Bart-base	**34.79**	**88.39**	3.03	2.93
Bart-large	32.80	88.23	2.99	3.05
T5-base	26.81	87.20	3.74	3.23
T5-large	27.14	87.38	**3.81**	**3.53**

Table 6: Generating changes resulting from a complex event (Task 3). See appendix Table 14 for the full results. The best scores are bolded.

Model	*Subset Acc*	*Hamming Score*	*macro F1*
Bart-base	64.6	**71.7**	61.2
T5-base (Enc Only)	59.4	66.1	50.3
T5-base (Enc-Dec)	**65.8**	67.3	**62.0**

Table 7: Results for identifying various Change Modes in Participants (Post Conditions) resulting from the endpoint of a complex action (Task 4).

6.6 Task 4: Identifying Types of Changes

We fine-tune base models on a multi-label (n=5) binary classification task and assign change mode labels, $(c_1, c_2, .., c_k)$, for an input context consisting of two sentences: story S, and (PI, c, ED) where c is a connector phrase. The value of c is "caused this:" when PR="agent," and "experienced this:" when PR="patient." The results from this classification are reported in Table 7 and consist of Subset accuracy, Hamming Score and Macro F1. The *Enc Only* models consists of a T5 encoder model with a classification head on top. The classification head consists of the following sequence of transformations: Dropout (p = 0.3) –> Linear(768x 512) –> tanh() –> Dropout (p = 0.3) –> Linear(512 x 5) –> sigmoid(). We also fine-tuned a pretrained T5 encoder-decoder model in a text-to-text multi-label RTE setting.

Discussion: These results indicate that while the fine-tuned models are good at generating change summaries, assigning the various change model labels is a challenging task for these LMs.

Model	Combined Acc./F1	Agent Acc./F1	Patient Acc./F1
Bart-base	82.7/76.8	**80.7/ 75.2**	84.2 /76.4
Bart-large	76.0/43.2	75.5/49.1	79.2/50.8
T5-base	82.6/76.5	80.1/73.2	84.2/77.4
T5-large	**83.0/77.4**	80.3/74.3	**84.5/78.1**

Table 8: Identifying Participant's involvement (Task 5). The best results are bolded. The different folds of the Bart-large models converge at different checkpoints resulting in lower average scores but the best scores for any fold are comparable to the Bart-base model.

6.7 Task 5: Assessing Participant Involvement

We turn the 5-point Likert scale for PI into a binary class: the first two options (unlikely to be involved) make up the negative class and the latter three (neutral to likely to be involved) are the positive class. We formulate participant involvement and enablement of changes as entailment: the story S is the premise and the hypothesis is framed as the P's involvement in the changes of state indicated by CS.[3] We fine-tuned models on all story annotations; only annotations where P's semantic role is "agent"; and only the annotations where P's semantic role is "patient." Results are in Table 8.

Discussion: While all the models are able to classify a participant's involvement and enablement of changes with high accuracy there is still room for improvement. Error analysis indicated models are not able to identify enablement when there are no state changes. This usually happens when the complex action is a hypothetical situation or the changes involve subtle cognition (discussed in Task 3). In T5 models, we noticed some errors contradicted the hypothesis statement; these may be due to the model's external knowledge from pre-training, but this requires further study.

7 Effect of Discourse Text on Models

We study the effect of discourse text on model generations of endpoint descriptions using the two story types we annotated: ROCStories and newswrire stories. ROCStories are simpler with short, concise and focused salient events, while newswire are more complex, containing more text not always salient to the complex action we annotated. We wish to answer the following questions: (1) How does training domain affect endpoint generation? (2) Are the endpoint generations more/less concise, varied, focused and factual for the story context? (3) Do models trained on one type of stories transfer their learnt knowledge to generate equally good endpoints for the other type?

[3]We formulate agent-based hypotheses as "What the actions of P caused was this: CS," and patient-based hypotheses as "What happened to P was this: CS."

Train-on	Model	Test-on	Len	Ext↓	Fact.	Sal.
ROC	Bart-base	ROC	5.6	.55	4.43	4.25
	T5-base	ROC	8.4	.55	4.43	4.24
	Bart-base	News	9.8	.39	3.98	3.59
	T5-base	News	23.8	.64	4.46	4.01
News	Bart-base	ROC	6.2	.89	4.55	4.09
	T5-base	ROC	9.0	.74	4.45	3.43
	Bart-base	News	11.5	.77	4.62	4.01
	T5-base	News	15.5	.75	4.59	3.95

Table 9: Comparing Endpoint generations of models trained on ROCStories and on Newswire stories. See appendix Table 15 for the full results.

We fine-tuned BART and T5 base models separately on ROCStories vs. Newswire, and evaluated them on test sets for both story types. We calculated human scores from the endpoint evaluation HIT. From Table 9, we observe the following: (1) ROCStories models generate shorter, more varied and abstract descriptions. (2) Newswire generations are longer and more extractive. (3) ROC-trained BART has significantly lower salience when tested on Newswire stories. News-trained BART does not suffer from poorer salience. News-trained T5 has lower salience when tested on ROCStories, while ROC-trained T5 does not result in significantly lower salience. (4) BART generations are less factual than T5, possibly because of higher abstractness (Dreyer et al., 2021). (5) ROC-trained T5 and News-trained BART obtain similar high scores for factuality and salience.

8 Conclusions

We have argued that a deeper understanding of complex events can be achieved by examining their cumulative outcomes, grounded as changes of state. By focusing on a specific participant in a complex event, and a broad notion of its semantic role, we developed a crowdsourcing protocol to obtain 7.7k annotations about complex events and participant state change across 4k stories. We validated 20% of the annotations, with high inter-annotator agreement. We have formulated five challenge tasks that stress model's understanding of story outcomes, state changes and complex event understanding. Our evaluations suggest that additional modeling advances are needed to achieve this understanding; we hope that our dataset spurs this future work.

9 Limitations

We acknowledge the following limitations of our approach:

- Though the documents we base our annotations on come from well known data sources, our efforts focus on more formal levels of written English. Generation and classification abilities can vary as the formality, style, or language change.
- Though our work is heavily grounded in interdisciplinary literature, we adopt a limited two-argument view of complex event participants: either they are an "agent" or a "patient." Expanding to other types, or finer-grained notions, of arguments requires more investigation.
- We use large, pre-trained language models in our experiments. While powerful, they can echo biases, either implicitly or explicitly. We do not attempt to control for these in this work.

Acknowledgements

We would like to thank the anonymous reviewers for their comments, questions, and suggestions. This material is based in part upon work supported by the National Science Foundation under Grant No. IIS-2024878. Some experiments were conducted on the UMBC HPCF, supported by the National Science Foundation under Grant No. CNS-1920079. This material is also based on research that is in part supported by the 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.

References

M. Bal. 1997. *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press.

Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. 2013. Generating Coherent Event Schemas at Scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731, Seattle, Washington, USA. Association for Computational Linguistics.

Mohaddeseh Bastan, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's Sentiment Prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John B. Black and Gordon H. Bower. 1980. Story understanding as problem-solving. *Poetics*, 9(1):223–250. Special Issue Story Comprehension.

Guido Boella, Rossana Damiano, and Leonardo Lesmo. 1999. Understanding narrative is like observing agents. *AAAI Technical Report FS-99-01*.

Peter Brooks. 1984. *Reading for the Plot: Design and Intention in Narrative*. Knopf Doubleday Publishing Group.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the Original: Fact Aware Neural Abstractive Summarization. *CoRR*, abs/1711.04434.

Tommaso Caselli and Oana Inel. 2018. Crowdsourcing StoryLines: Harnessing the Crowd for Causal Relation Annotation. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 44–54, Santa Fe, New Mexico, U.S.A. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2016. The Storyline Annotation and Representation Scheme (StaR): A Proposal. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 67–72, Austin, Texas. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

T. A. Dijk and W. Kintsch. 1983. Strategies of Discourse Comprehension. In *Psychology*.

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. In *Language*, volume 67, pages 547–619, USA. Linguistic Society of America.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2021. Analyzing the Abstractiveness-Factuality Tradeoff With Nonlinear Abstractiveness Constraints. *CoRR*, abs/2108.02859.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To Test Machine Comprehension, Start by Defining Comprehension. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 7839–7859, Online. Association for Computational Linguistics.

Joshua Eisenberg and Mark Finlayson. 2017. A Simpler and More Generalizable Story Detector using Verb and Character Features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2708–2715, Copenhagen, Denmark. Association for Computational Linguistics.

Shen Gao, Xiuying Chen, Piji Li, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2019. How to Write Summaries with Patterns? Learning towards Abstractive Summarization through Prototype Editing. 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 3741–3751, Hong Kong, China. Association for Computational Linguistics.

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A Corpus for Extracting Event Hierarchies from News Stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3678–3683, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30. ACM.

Arthur C. Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101 3:371–95.

H. P. Grice. 1975. Logic and Conversation. *Syntax and Semantics*, 3:41–58.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. ESTER: A Machine Reading Comprehension Dataset for Event Semantic Relation Reasoning. In *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David Herman. 2002. Problems and Possibilities of Narrative. In *Story Logic*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *CoRR*, abs/1907.10529.

Aaron N. Kaplan and Lenhart K. Schubert. 2000. A computational model of belief. *Artificial Intelligence*, 120(1):119–160.

Walter Kintsch and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85:363–394.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Heeyoung Kwon, Mahnaz Koupaee, Pratyush Singh, Gargi Sawhney, Anmol Shukla, Keerthi Kumar Kallur, Nathanael Chambers, and Niranjan Balasubramanian. 2020. Modeling Preconditions in Text with a Crowd-sourced Dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3818–3828, Online. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. The Future is not One-dimensional: Complex Event Schema Induction by Graph Modeling for Event Prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5203–5215, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

G. Liveley. 2019. *Narratology*. Classics in theory. Oxford University Press.

Peter LoBue and Alexander Yates. 2011. Types of Common-Sense Knowledge Needed for Recognizing Textual Entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101.

Donata Marasini, Piero Quatto, and Enrico Ripamonti. 2016. Assessing the inter-rater agreement for ordinal data through weighted indexes. *Statistical Methods in Medical Research*, 25:2611 – 2633.

Michael Mateas and Phoebe Sengers. 1999. Narrative Intelligence. *AAAI Technical Report FS-99-01*.

Ted Matherly. 2018. A Panel For Lemons? Positivity bias, reputation systems and data quality on MTurk. *European Journal of Marketing*, 53.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating Causality in the TempEval-3 Corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan. The COLING 2016 Organizing Committee.

Daniel G Morrow, Gordon H Bower, and Steven L Greenspan. 1989. Updating situation models during narrative comprehension. *Journal of Memory and Language*, 28(3):292–312.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative Theory for Computational Narrative Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

James Pustejovsky, Robert Ingria, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *In New Directions in Question Answering*.

Zheng Qi, Elior Sulem, Haoyu Wang, Xiaodong Yu, and Dan Roth. 2022. Capturing the Content of a Document through Complex Event Identification. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 331–340, Seattle, Washington. Association for Computational Linguistics.

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.

Sebastian Raschka. 2018. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *The Journal of Open Source Software*, 3(24).

Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script Induction as Language Modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1681–1686, Lisbon, Portugal. Association for Computational Linguistics.

Marie-Laure Ryan. 1991. *Possible worlds, artificial intelligence, and narrative theory*. Indiana University Press, Bloomington.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Lenhart K. Schubert and Chung Hee Hwang. 2000. *Episodic Logic Meets Little Red Riding Hood: A Comprehensive Natural Representation for Language Understanding*, page 111–174. MIT Press, Cambridge, MA, USA.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *CoRR*, abs/1704.04368.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly Supervised Subevent Knowledge Acquisition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5345–5356, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

RA Zwaan and GA Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162—185.

A Additional Details on Data Preparation

In this section, we expand on data processing described in §4.1.

Document Selection From the Annotated New York Times (ANYT) newswire articles, we found that stories from the Financial, National and Foreign desks contained the type of complex events that were most reliable to annotate: those with focused discourse text that required less external, societal, or cultural knowledge to understand the story. We did not specifically target obituaries as they could lead to less varied endpoint and cumulative state changes. The ROCStories were randomly sampled, and we subsampled stories from ESTER that had `subevent` annotations in that dataset.

Participant Identification We used spanBERT (Joshi et al., 2019) to resolve coreferent mentions in the text, and selected the largest cluster of mentions. To find clusters containing a valid participant, we selected the shortest text span from all the mentions in the cluster making sure that it is atleast 3 characters long and matched it with the names database published by the SSA. This ensured that the "participant or prop" we selected is a person, place, group or organization. In ROCStories and ESTER, the largest cluster is always a person, place or group and did not require this name filtering.

Continuant Story Lines We selected the first few lines containing approximately 100 tokens, which resulted in stories similar in length to previous work (Han et al., 2021; Glavaš et al., 2014; O'Gorman et al., 2016). We highlighted mentions of the participant to outline a "continuant" storyline, i.e., a set of related events that lead to a coherent story involving the participant. Focusing on the events in a continuant story line helps an annotator observe a complex action and its effects. By identifying and highlighting a single participant we limit the scope of possible valid endpoints an annotator might consider. Assigning a semantic role to the participant, of an agent, or a patient, helps cue the annotator to identify a participant's role in the complex action, and the changes resulting from it.

B Additional Details on Crowd Annotation

B.1 Worker Qualifications

We did not use requester generated qualification tests to filter out workers because we target the understanding of everyday reported events, not domain or expert-level matter. However, we used community standard quality criteria, such as requiring a 98% or greater HIT acceptance rate and the completion of 1000 approved HITs. In addition, we required the worker's stated location to be in the USA, UK, Canada, Australia, or New Zealand. Given the language-dependent semantic phenomena we pursue in this work, this location requirement was used for avoiding language-based artifacts. While qualification tests can filter for spam, initial misunderstandings could exclude capable workers who benefit from additional feedback. By providing positive and constructive feedback to ensure workers understood the task, we were able to retain workers who improved over time and provided quality annotations, a requirement for any crowdsourced task.

B.2 Annotation HIT Streamlining

Our initial development tested selection of textspans vs. free form text and noticed workers preferred one over the other for some of the steps. To reduce annotation time, we refined the HIT to prime workers to hone in on the salient information in the story, provided functionality that allowed for a quick highlight and paste of relevant text when needed. We encouraged free form text in steps 1 and 2., an easy to fill in template for step 3, and highlight and paste for step 4.

Despite instructions to be concise, early annotations suggested that some workers would try to include as much information as possible into the free form text fields, resulting in lengthy descriptions that provided too much detail (e.g., going beyond immediate outcomes, or providing explanation/justification for why those changes happened). To address this issue, we implemented two-tiered length limitations on the free form text. The first tier was a "soft constraint": if, e.g., a worker typed in a endpoint greater than 8 words, they were prompted to consider revising, but they did not have to. The second tier was a "hard constraint": if, e..g, the endpoint was greater than 15 words, they were prevented from submitting until they rephrased and satisfied the hard constraint limit. These limits were set based on the examination of early annotations.

In addition, in each HIT batch, we included a mix of the lengthier Newswire and short ROCStory texts to reduce the monotony of annotation. From the alpha run annotation times, and internal annotation timing, we estimated the average annotation time for a HIT completion to be under 2 minutes. The bulk of annotations for our HITs were completed within 5 minutes, with a median and mean of approximately 2.5 minutes.

B.3 Worker Training and Pay

Our HITs were priced to target an hourly pay of $10-$12. We carefully tracked and analyzed the user response times across pilot runs to arrive at the HIT pricing. For each worker, we carefully examined the first 10-30 annotations to check task understanding, providing feedback and a bonus as appropriate to compensate for the time spent on communication. We initially had an additional 26 crowd workers who attempted the HIT, but we removed their annotations from the dataset for obvious bad-faith efforts (10 workers) and for benefit-of-the-doubt good faith efforts but where the workers (16 workers) did not follow instructions even with repeated feedback. Anyone construed to have completed the annotation in good faith was paid, even when their responses were not included in our dataset.

B.4 Annotation Quality Checking

Our cursory visual check of the annotation and an automatic lexical check of a list of novel unigrams that are not part of the story ensured the annotation content was focused on the story. We gave iterative feedback to workers not following task instructions and excluded them with a qualification type when there was no improvement.

Additional Data Analysis In Fig. 4 we show the distribution of crowd workers for our annotation effort.

B.5 Annotation HITs

The HIT for acquiring story annotations from crowd workers displays two different views based on participant's semantic role in the story. General and specific instructions for Step 1-3 are different in two views.

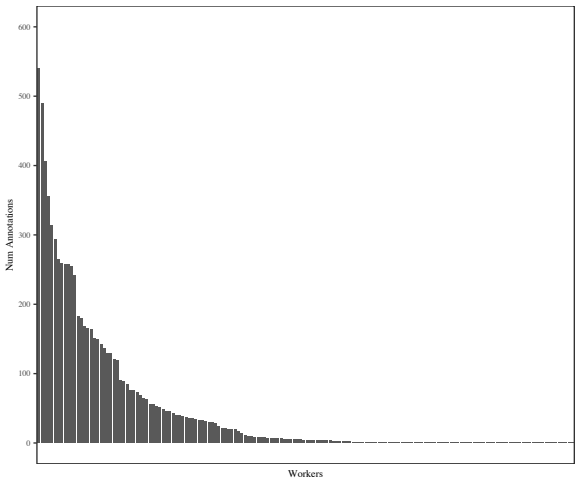

Figure 4: Distribution of story annotations completed by workers

C Quality Assessment of Crowd Annotations

In the initial phases of data collection the first two authors of this paper evaluated both the agent and patient views of 50 random stories (= 100 annotations) to ensure annotator responsiveness and quality. After collecting all the data, 3 crowd workers evaluated a random 1545 annotations and an expert evaluated a random 100 of these annotations (equal agent and patient views) for comparison. Table 10 lists the inter-rater reliability (IRR) measured using weighted Fleiss's Kappa (Marasini et al., 2016) with the weighting scheme used by (Bastan et al., 2020), which penalises each dissimilar class by an amount based on the distance between classes (e.g., an item with responses of "very likely" and "very unlikely" will be penalized more heavily than with responses of "very likely" and "somewhat likely").

Our evaluation consisted of 4 validation HITs, where 3 crowd workers rated the various annotation steps (see Appendix C.2). The results from this evaluation are listed in Table 10. The IRR scores suggest substantial-to-high agreement. Notably, these demonstrate that we can obtain high quality process and change of state summaries, endpoint descriptions and enabling sub-event factors.

C.1 Evaluation Set Curation

From the 1545 validated annotations we selected annotations where the average score of the crowd workers for each of the 4 validation HITs is at least 3.0. This curation resulted in 1196 carefully produced annotations for a given story. The test data set is made up of these curated annotations and the training data set is made up of the remaining validated and unvalidated annotations.

Evaluations	*Crowd*	*C+E*	*Experts*
1. Process Sum.	0.81	0.81	0.90
2. Endpoint Desc.	0.81	0.80	0.89
3. Change Sum.	0.81	0.80	0.76
4. Change Modes	0.74	0.78	0.82
5. Factors' Salience	0.77	0.85	0.84

Table 10: Inter-rater Agreement scores using weighted Fleiss's Kappa (Marasini et al., 2016). C+E shows the IRR for the crowd and expert on 100 annotations. See Appendix C.2 for the various evaluations and what we looked for in the evaluation.

C.2 Validation HITs

The various annotation steps are validated using 4 HITs. 3 workers evaluate the following using a 1-5 Likert scale with the options: Strongly Disagree, Somewhat Disagree, Neutral, Somewhat Agree and Strongly Agree.

1. Whether the Process Summary is a valid high level summary of the story.
2. Whether the endpoint description describes a valid endpoint for the complex action in the story.
3. Whether the change summary describes changes that happened as result of the complex event described in the story.
4. Whether the categorization of changes into the five change modes is consistent with the changes inferred from the story.
5. Whether the factors are salient to the complex event's endpoint.

D HITs for Human Evaluation of Generated Summaries and Factors

Here, we show sample UIs for the human evaluation we performed in §6.2 for reference, baseline and model-generated process (Fig. 17), endpoint (Fig. 18), change summaries (Fig. 19) and factors (Fig. 20) . These HITs are used to evaluate the following aspects of a summary using a 1-5 likert scale with the options Strongly Disagree, Somewhat Disagree, Neutral, Somewhat Agree and Strongly Agree.

1) *Abstractness (Task 1):* Whether the summary is a brief, high level, abstract description that faithfully captures the complex action in the story.

2) *Validity (Tasks 2 and 3):* Whether the summary is a valid ending for the situation described in

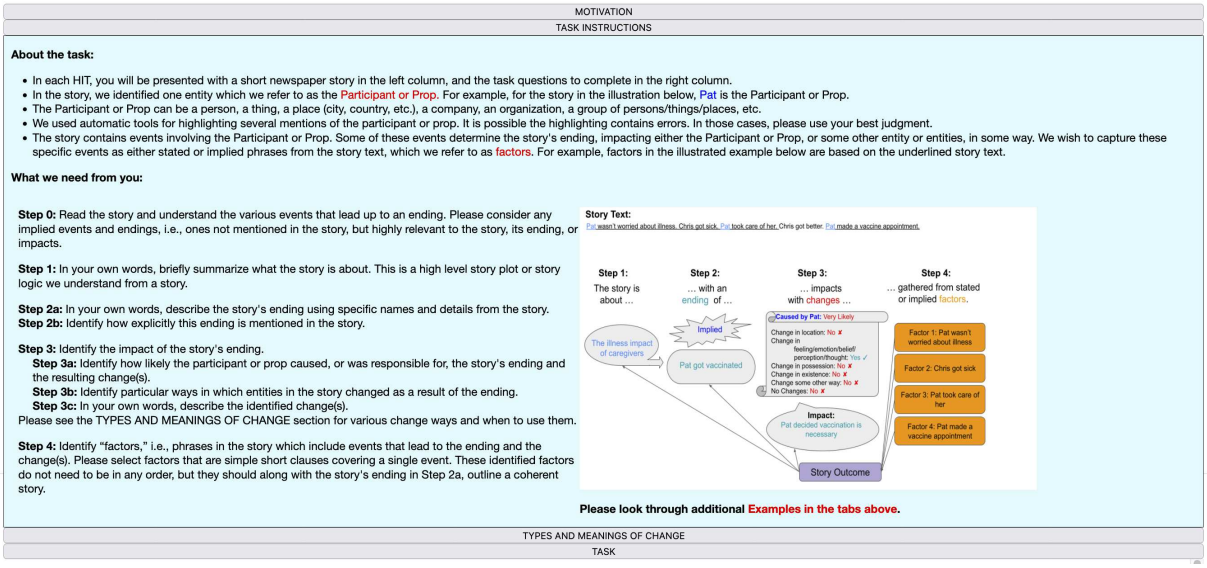

Figure 5: Instructions provided for the Agent view of the annotation HIT. The distinguishing aspect that makes it the Agent view is in step 3, where changes are attributable to what the participant or prop caused. In 7, 8, 10 and 12 we show the interface for each of the steps.

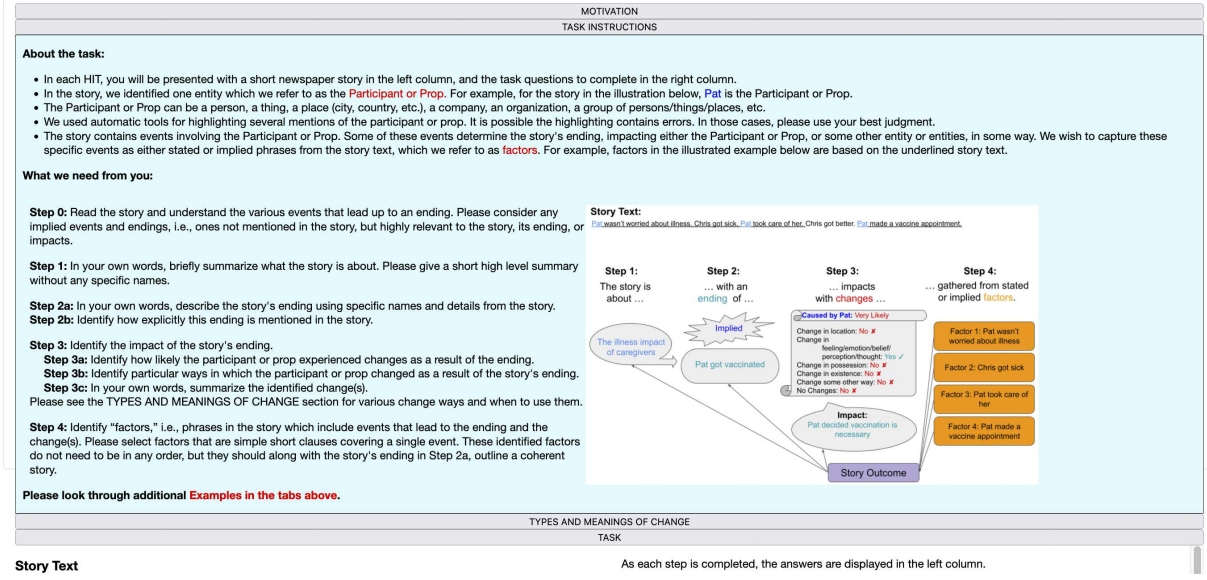

Figure 6: Instructions provided for the Patient view of the annotation HIT. The distinguishing aspect that makes it the Patient view is in step 3, where we ask about changes the participant or prop likely experienced. In 7, 9, 11 and 12 we show the interface for each of the steps.

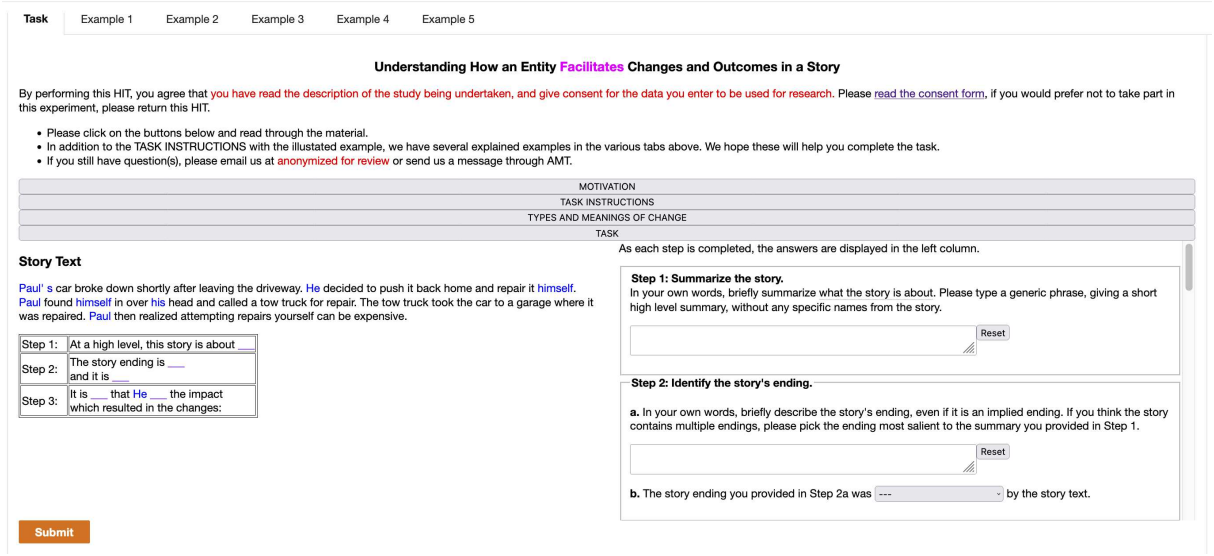

Figure 7: Steps 1 and 2 of the Agent view of the annotation HIT. The Patient view for these steps is similar except for the title of the HIT.

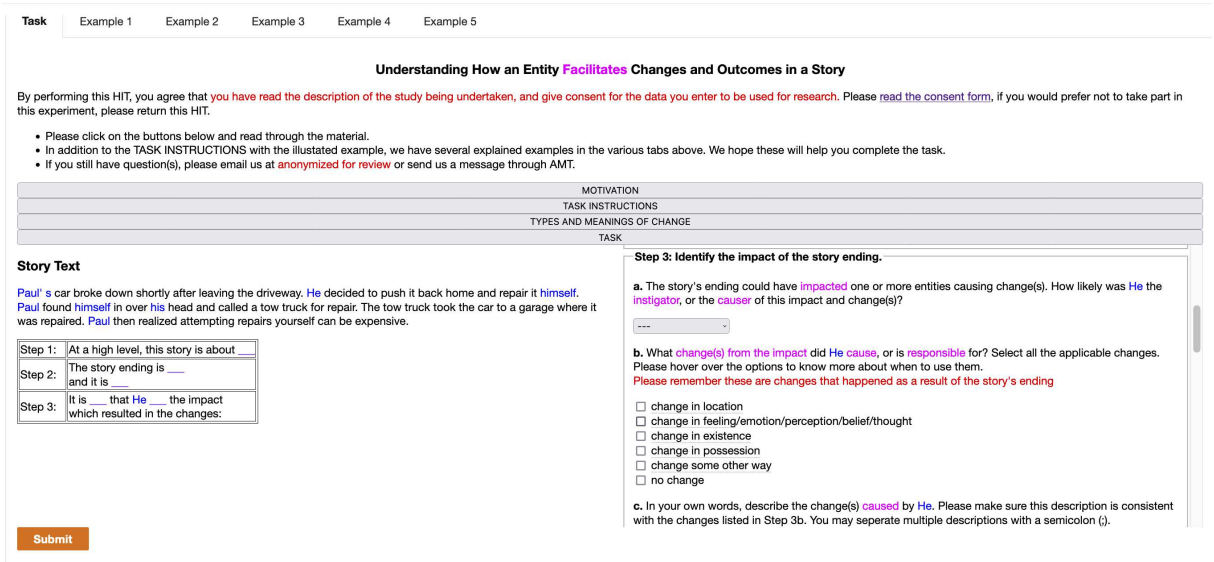

Figure 8: Steps 3a & 3b of the Agent view of the annotation HIT.

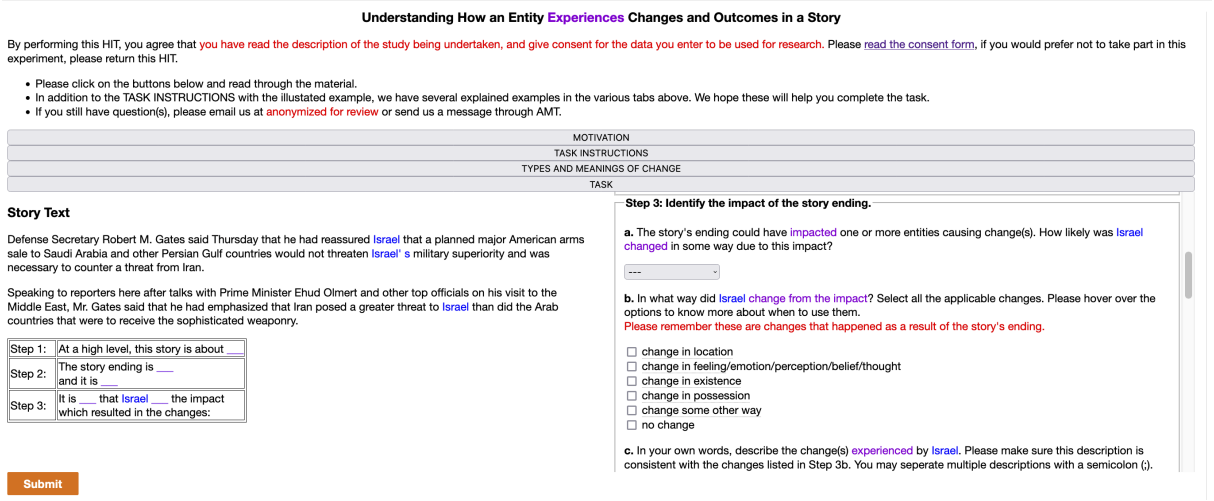

Figure 9: Steps 3a & 3b of the Patient view of the annotation HIT.

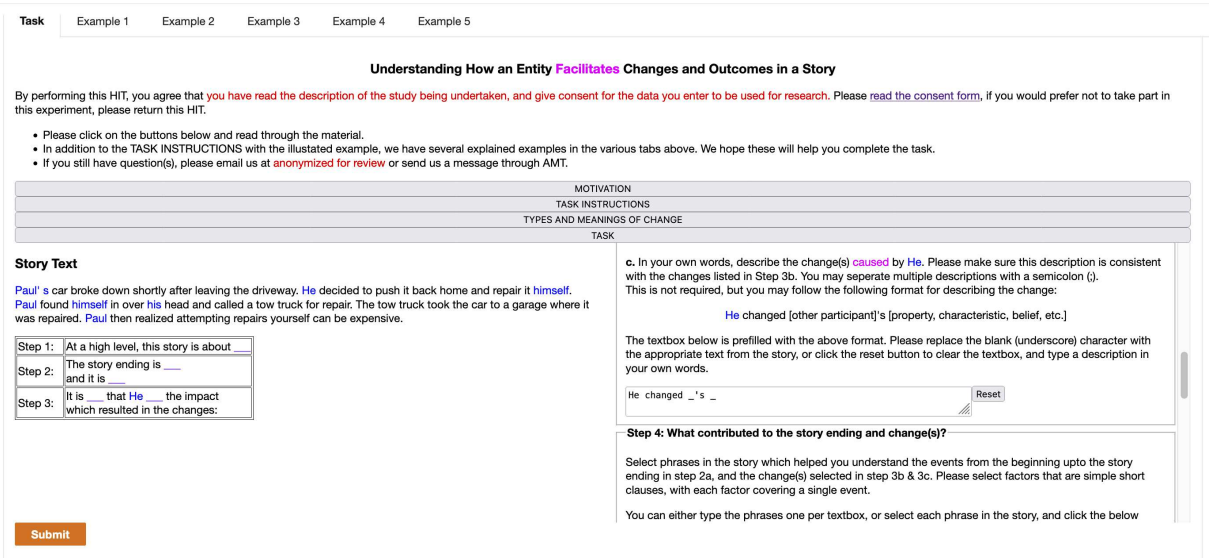

Figure 10: Step 3c of the Agent view of the annotation HIT.

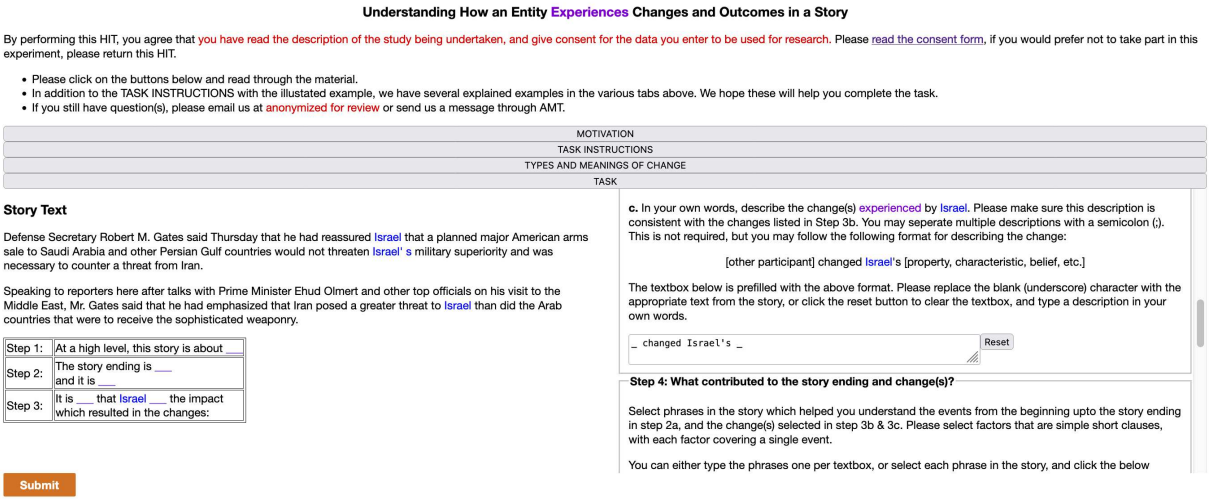

Figure 11: Step 3c of the Patient view of the annotation HIT.

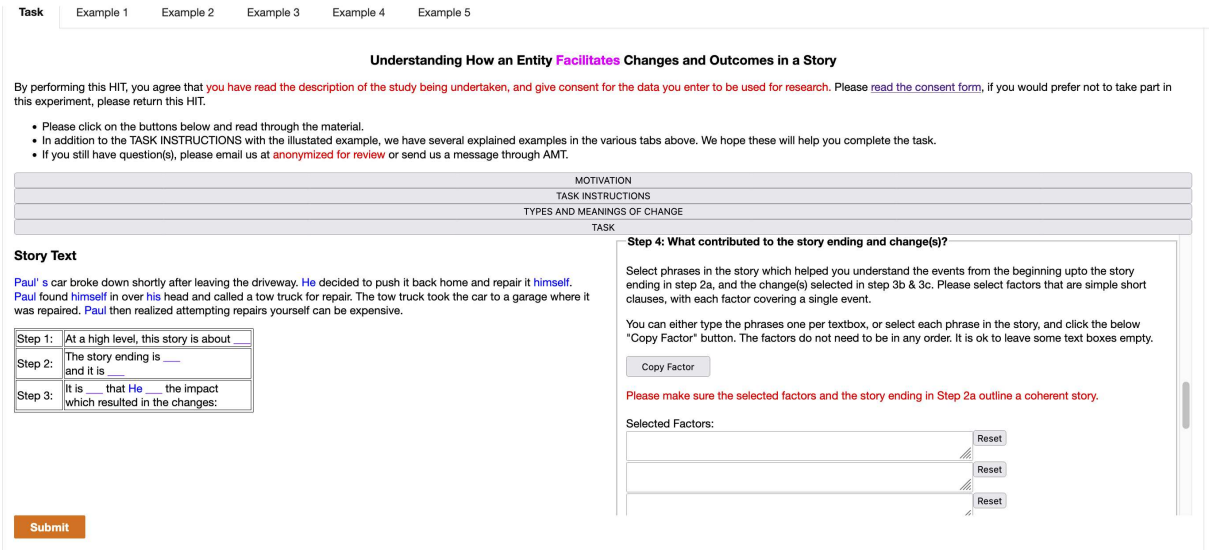

Figure 12: Step 4 of the Agent view of the annotation HIT. The Patient view for this step is similar except for the title of the HIT.

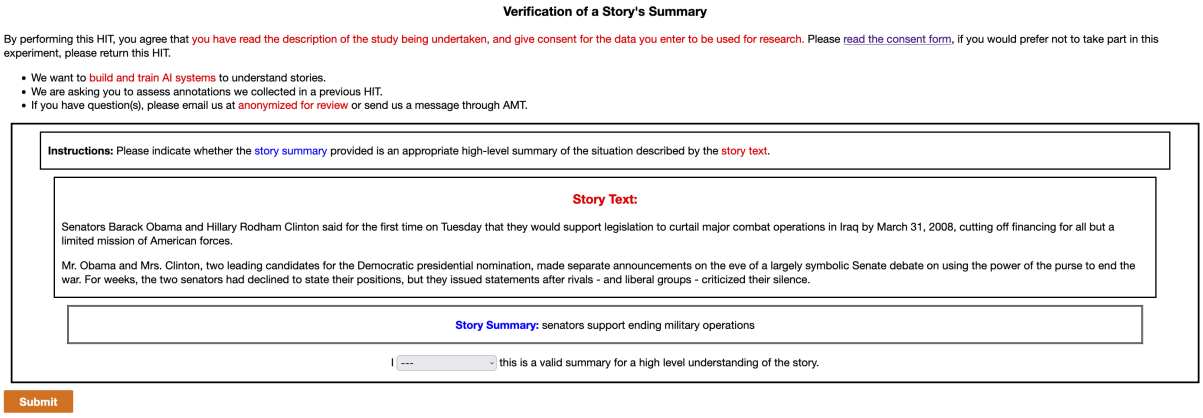

Figure 13: HIT for Validating Process Summaries

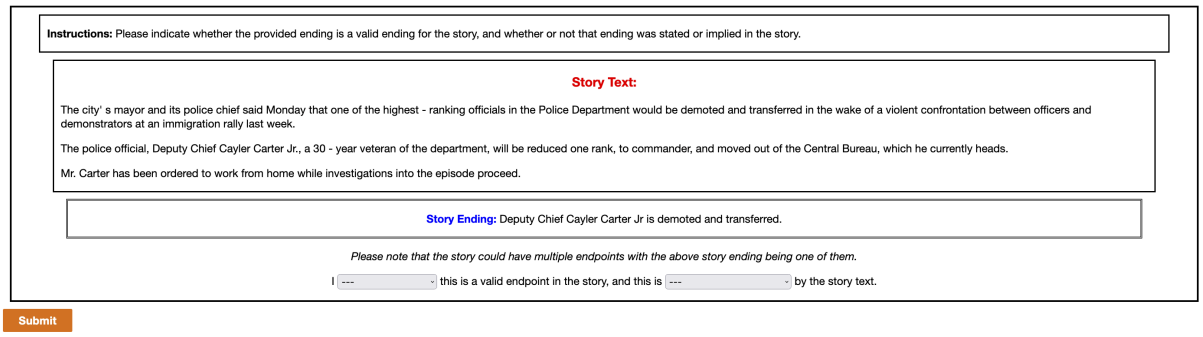

Figure 14: HIT for Validating Endpoint Descriptions

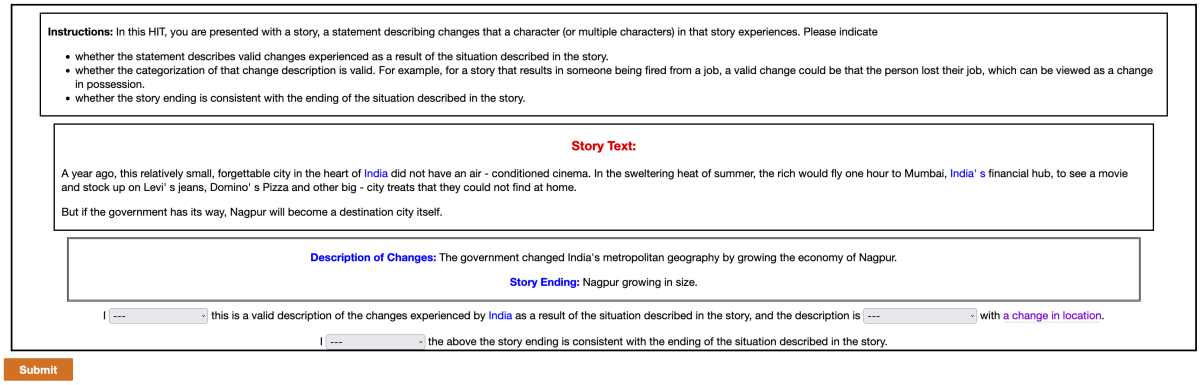

Figure 15: HIT for Validating Change Summaries

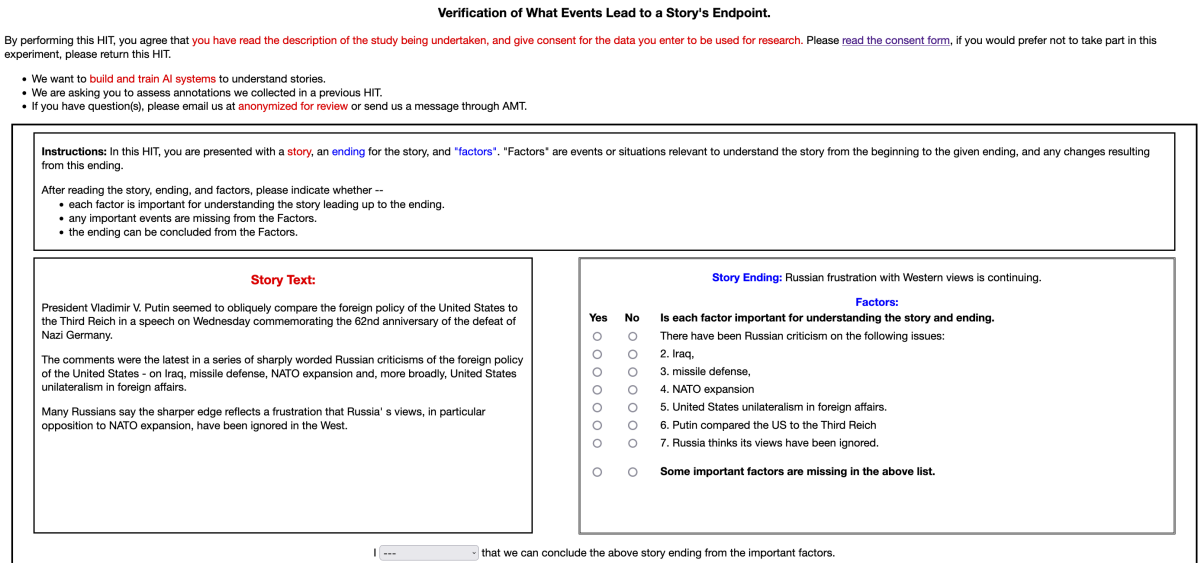

Figure 16: HIT for Validating Factors

the story for task 2. In task 2, when evaluating factors, we check whether each factor contains story related information. For task 3, we check whether the summary mentions a change that is consistent with the details in the story. A blank summary is valid, if the story does not contain any changes.

3) *Salience (Tasks 2 and 3):* Whether the summary is a valid ending for the situation described in the story for task 2. In task 2, when evaluating factors, we check whether each factor is necessary for understanding the situation. For task 3, we check whether the summary mentions a change that is consistent with the details in the story. When a story does not contain any changes, the annotation of no-changes is considered salient.

E Model Training

To perform early development experiments, we used 5-fold cross validation for 2 epochs. We found that while the evaluation loss plateaued after training for an epoch, recall improved with further training for another epoch. Classification results are the average of the 5-fold cross validation after 2 epochs for T5 and 1 epoch for BART for the single label classification in Task 5. The Multi-label binary classification for Task 4 was trained for 5 epochs for the BART models, 10 epochs for the T5 Encoder Only and the Encoder-Decoder model. For the summarization tasks, we trained models on the entire training set without subsequent hyperparameter tuning for 2 epochs for all tasks.

F Expanded Results

In this section we present expanded results from §6. In addition to average number of tokens in a summary or factor (*Len*) and percentage of extractive trigrams from the story text (*Ext*), ROUGE-L (longest word sequences) and BertScore that were reported in the paper, we use the ROUGE scores based on unigrams (ROUGE-1) and bigrams (ROUGE-2), corpus and sentence level BLEU, and METEOR. Unlike the lexical- and ontological-based metrics of ROUGE, BLEU and METEOR, BertScore aims to provide a modern, embedding-based approach for handling semantic equivalence/similarity even when the texts being compared have different surface forms (e.g., different words are used). These automated scores are calculated using all the test data where as the human evaluation scores are for the 50 randomly selected data items for evaluation. The automated scores of ROUGE, BLEU, METEOR and BertScore are not reported for the reference summaries, as these summaries were used as gold references when calculating automated scores for the baseline and model generations. For the human evaluation we report the IAA by crowd workers. The difference in IAA for the crowd + expert scores and the crowd scores was < 0.1. The metrics reported on the left of the double line are calculated for all the test data.The best value for each score category are bolded and those that are significantly higher than all other values are marked with a *.

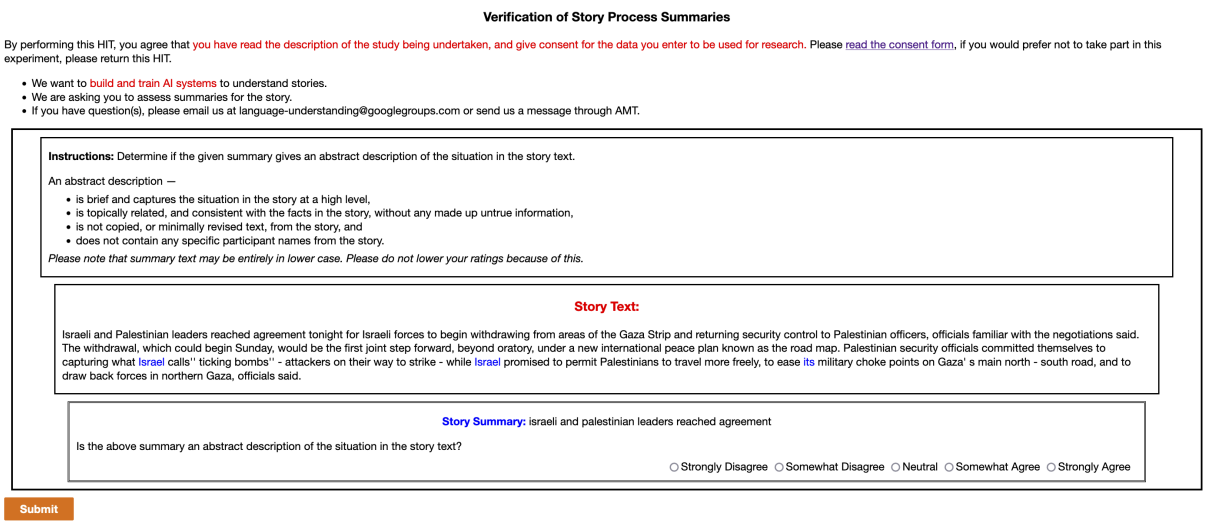

Figure 17: HIT for Evaluating Reference, Baseline and Generated Process Summaries

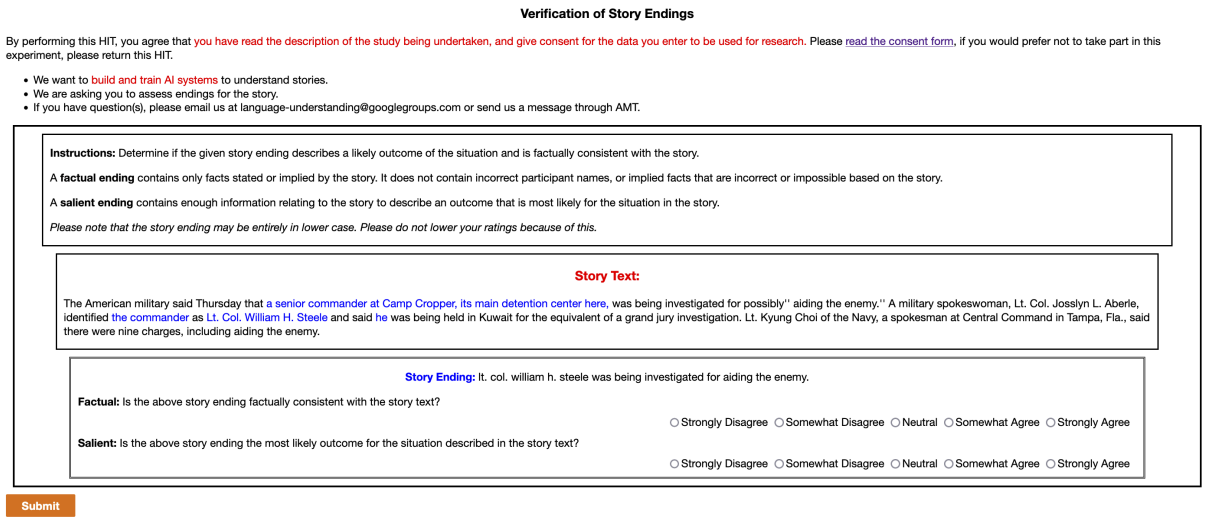

Figure 18: HIT for Evaluating Reference, Baseline and Generated Endpoint Summaries

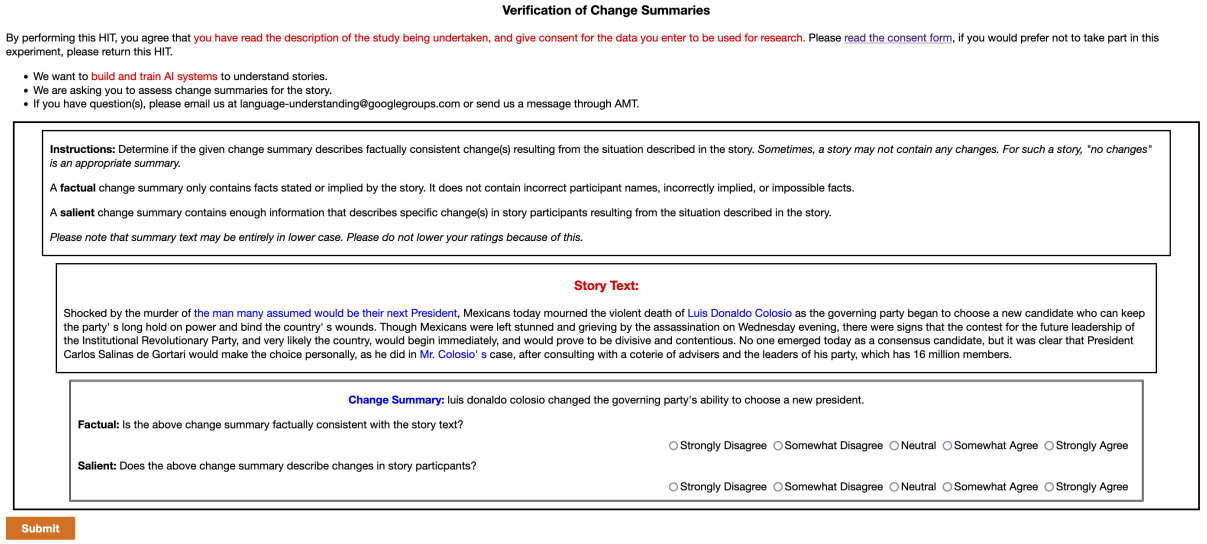

Figure 19: HIT for Evaluating Reference and Generated Change Summaries

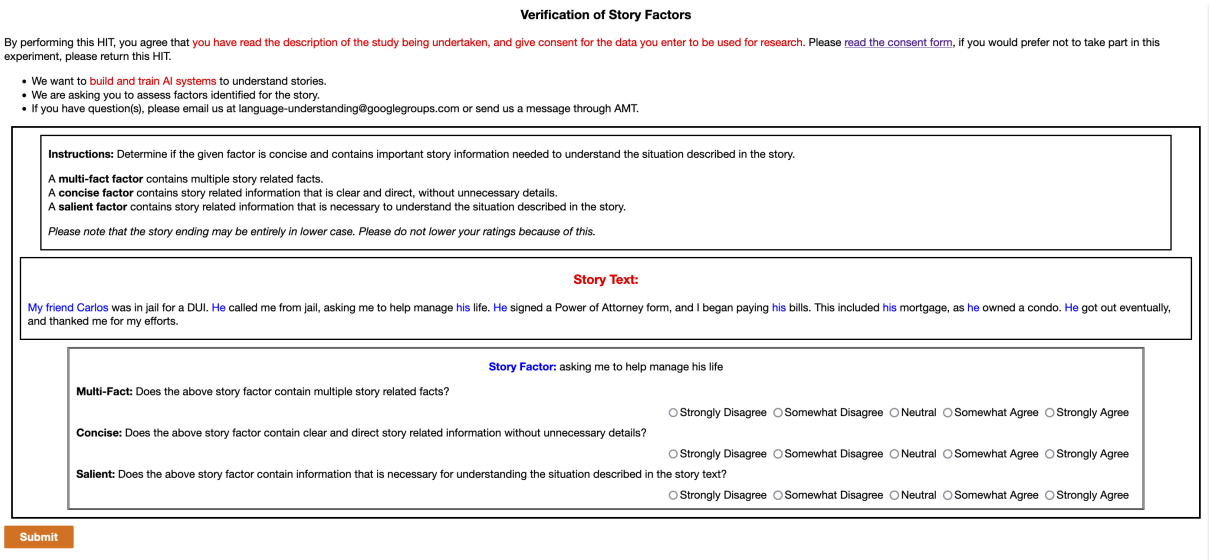

Figure 20: HIT for Evaluating Reference and Generated Factors

				Rouge			MET-	BLEU		Bert	Abstractness		
	Model	Len	Ext↓	1	2	L	EOR	sent.	corpus	Score	AvgLS	IAA	%Abs
	Reference	3.6	.13*				-				**3.57**∗	.82	72
	About P	1.7	.27	10.55	4.22	10.43	0.07	0.35	2.70	83.86	2.37	.48	16
Story	Bart-base	4.0	.46	22.43	8.34	21.43	15.26	**7.89**	4.27	86.70	2.77	.86	42
	T5-base	10	.60	21.14	7.48	19.50	**18.43**	5.55	2.81	85.99	2.32	.72	28
	T5-large	6.9	.63	21.73	8.47	20.30	17.59	7.78	4.16	86.15	2.13	.77	20
Fact.	Bart-base	4.2	.33	**25.17**	**9.05**	**23.81**	17.61	6.73	**4.89**	**87.66**	3.22	.76	54
	T5-base	9.9	.56	19.94	6.39	18.29	17.47	4.19	2.63	86.10	2.73	.69	36

Table 11: Generating Process Summaries for stories and factors+endpoint (Task 1). The best scores are bolded. * indicates the Reference value for Abstractness is significantly higher than other values in the column with a p value between 0.001 - 0.0001 (except for Bart-base trained on Factors where the p value is 0.13). *Len* value for Reference is considered the best as the baseline value is not meaningful. The column AvgLS is the average of 3 crowd worker 1-5 Likert scores, and the column % Abs is percentage of instances with a score ≥ 3.

				Rouge			MET-	BLEU		Bert	Factuality			Salience		
	Model	Tok	Ext↓	1	2	L	EOR	sent.	corpus	Score	AvgLS	IAA	%Abs	AvgLS	IAA	%Abs
	Reference	7.9	.27				-				4.15	.87	80	3.46	.76	68
	Last sent.	23.3	.79	24.08	12.33	21.82	**24.62**	0.58	6.59	85.60	4.49	.82	90	3.35	.82	62
Story	Bart-base	11	.72	27.83	13.05	25.43	22.88	**11.12**	**8.31**	87.61	4.66	.73	96	3.97	.69	80
	Bart-large	10.3	.63	27.00	12.51	24.74	21.44	10.82	8.30	87.62	4.59	.78	96	3.81	.72	74
	T5-base	13.6	.70	26.66	11.79	24.07	22.69	9.79	6.62	87.19	**4.71***	.75	96	4.03	76	82
	T5-large	12.9	.67	**28.36**	**13.15**	**25.71**	24.31	10.93	7.83	87.54	**4.71***	.77	100	**4.23***	72	90
Fact.	Bart-base	7.3	.49	26.26	10.86	24.09	18.28	9.15	7.32	**87.93**	4.11	.77	80	3.28	.76	52
	T5-base	10.4	.47	24.74	9.52	22.01	18.53	7.31	5.69	87.07	3.99	.72	82	3.02	69	44

Table 12: Generating Endpoints for stories and factors (Task 2a). The best scores are bolded. * indicates the value is significantly higher than the Reference with a p value of .002 for Factuality and .0008 for Salience. The column AvgLS is the average of 3 crowd worker 1-5 Likert scores, and the column % Abs is percentage of instances with a score ≥ 3.

	Factors		Rouge			MET-	BLEU		Bert	Brevity			Factuality			Salience		
Model	Num	Len	1	2	L	EOR	sent.	corpus	Score	AvgLS	IAA	%Abs	AvgLS	IAA	%Abs	AvgLS	IAA	%Abs
Reference	3.6	8.3								**3.35**	.73	90	3.25	.67	52	3.04	.73	42
Bart-base	3.5	14.0	49.90	37.35	45.28	43.04	0.42	25.18	88.06	2.46	.62	.60	3.49	.69	59	3.57	.76	63
Bart-large	3.6	13.7	50.51	38.11	45.98	43.63	**0.44**	25.32	88.10	2.88	.65	67	3.2	.68	50	3.31	.73	52
T5-base	2.6	19.6	47.89	35.86	43.31	44.12	0.38	24.61	87.74	1.77	.66	38	3.69	.76	68	**4.01***	.79	79
T5-large	3.7	13.9	**52.26**	**40.57**	**47.96**	**46.83**	0.40	**26.8**	**88.44**	2.15	.65	48	**3.80***	.70	71	3.96	.77	78

Table 13: Generating Factors from stories and their endpoints (Task 2b). The best scores are bolded. The * for factuality and salience indicates the value is significantly higher than the Reference with a p value of .0001. The Reference value for Brevity is significantly higher than all values in the column with a p value of .0001. The column AvgLS is the average of 3 crowd workers' 1-5 Likert scores, and the column % Abs is percentage of instances with a score ≥ 3.

Model	Rouge			MET-	BLEU		Bert-	Factuality			Salience		
	1	2	L	EOR	sent.	corpus	Score	AvgLS	IAA	%Abs	AvgLS	IAA	%Abs
Reference	-							3.36	.78	60	3.32	.76	64
Bart-base	**39.56**	**23.66**	**34.79**	**28.68**	**7.97**	5.29	**88.39**	3.03	.69	48	2.93	.71	50
Bart-large	39.28	20.68	32.80	27.87	7.10	6.91	88.23	2.99	.72	44	3.05	.74	48
T5-base	32.82	14.74	26.81	24.62	5.91	6.73	87.20	3.74	.75	52	3.23	.66	66
T5-large	34.40	15.53	27.14	24.65	6.11	**7.07**	87.38	**3.81**	.70	70	**3.53**	.70	72

Table 14: Generating changes resulting from a complex event (Task 3). The best scores are bolded. The column AvgLS is the average of 3 crowd workers' 1-5 Likert scores, and the column % Abs is percentage of instances with a score ≥ 3.

Trained-on	Model	Test-on	Len	Ext	Rouge			MET-	BLEU		Bert	Factu-	Sali-
					1	2	L	EOR	Sent.	corpus	Score	uality	ence
ROC	Bart-base	ROC	5.6	.55	42.54	21.13	39.16	31.20	18.54	12.37	91.46	4.43	4.25
	T5-base	ROC	8.4	.55	43.13	20.66	38.53	33.35	16.56	9.77	90.83	4.43	4.24
	Bart-base	News	9.8	.39	22.53	8.25	20.49	16.27	6.46	4.97	87.17	3.98	3.59
	T5-base	News	23.8	.64	23.60	10.03	20.60	24.54	6.99	4.64	86.44	4.46	4.01
News	Bart-base	ROC	6.2	.89	39.39	18.80	36.41	29.64	15.55	10.95	90.46	4.55	4.09
	T5-base	ROC	9.0	.74	35.42	14.41	31.39	26.94	11.75	7.01	90.26	4.45	3.43
	Bart-base	News	11.5	.77	25.28	11.58	23.12	20.77	10.16	7.78	86.95	4.62	4.01
	T5-base	News	15.5	.75	22.94	9.69	20.51	20.22	8.05	5.48	86.51	4.59	3.95

Table 15: Comparing Endpoint generations of models trained on ROCStories and models trained on Newswire stories

G Computing Infrastructure & Processing information

Infrastructure We trained our models on a single RTX 8000 with 48GB of GPU memory. Approximate run time was 1 hour.

Model Parameters In addition to the number of parameters in each of the models we consider (e.g., Bart-Base, T5-base, T5-large), each of our fine-tuned classification models has a separate classifier layer. This layer takes in a D dimension embedding from the encoder and uses a single linear layer to compute K dimensional logits (therefore, an additional DxK parameters). The value of D will depend on the model, and the value of K will depend on the number of label types that could be predicted. For the generation tasks, we do not add any additional parameters to the models.

Hyperparameters For all experiments we used AdamW (Loshchilov and Hutter, 2017) optimizer, a learning rate of $10 - 4$, a weight decay of $10^{-}4$ and a random seed of 11. We applied manual tuning and tried various learning rates from .001 to .00001 as suggested for the BART and T5 models. For the generation we used Top-K sampling with a beam size of 2. These parameters worked well for all the models and this was selected based on accuracy for the classification tasks and Rouge scores for the summarization tasks.

Results statistics With our 5-fold cross-validation, the automated metrics for summarization and the accuracy/F1 values for classification varied by less than 3 percent.