CoRE: Condition-based Reasoning for Identifying Outcome Variance in Complex Events

Sai Vallurupalli, Francis Ferraro
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250 USA
{kolli,ferraro}@umbc.edu

Abstract

Knowing which latent conditions lead to a particular outcome is useful for critically examining claims made about complex event outcomes. Identifying implied conditions and examining their influence on an outcome is challenging. We handle this by combining and augmenting annotations from two existing datasets consisting of goals and states, and explore the influence of conditions through our research questions and *Condition-based Reasoning* tasks. We examine open and closed LLMs of varying sizes and intent-alignment on our reasoning tasks and find that conditions are useful when not all context is available. Models differ widely in their ability to generate and identify *outcome-variant* conditions which affects their performance on outcome validation when conditions are used to replace missing context. Larger models like GPT-4o, are more cautious in such less constrained situations.

1 Introduction

Knowing which conditions influence a goal's outcome is useful for understanding and planning goal-directed actions seen in complex events (Csibra and Gergely, 2007). Consider the following 5-sentence short story, also shown in Fig. 1:

> Sam can't sleep at night. Sam is afraid of the monsters under the bed. Dad tells Sam there is no such thing as monsters but it doesn't help. So, dad give Sam a blanket and tells Sam that it's a magic blanket. Sam believes the blanket protects against monsters.

While one can reasonably infer that Sam's goal (to sleep at night overcoming fears about monsters) was achieved *in this story*, how stable is this outcome? What states, or conditions, relating to Sam or the situation help support the outcome, and if those were changed, would Sam's goal still be achieved? Some of these conditions, like Sam trusting his/her dad or Sam being a small child, are relevant to the goal, with a high likelihood of influencing the outcome, while other conditions that might be true are irrelevant to the outcome—like Sam being male. Among relevant conditions, Sam trusting his/her dad directly influences the outcome and is **outcome-variant** because if Sam did not trust his/her dad the outcome is less likely to be true. Alternatively, Sam being a small child does not directly influence the outcome; it is **outcome-invariant** as the contrastive condition is unlikely to change the outcome. In this paper, we examine the interplay between state conditions and goal outcomes, and in particular, **we demonstrate how to use, and generate outcome-variant counterfactual conditions to effectively reason about whether a story outcome is true (or false) across altered narratives.**

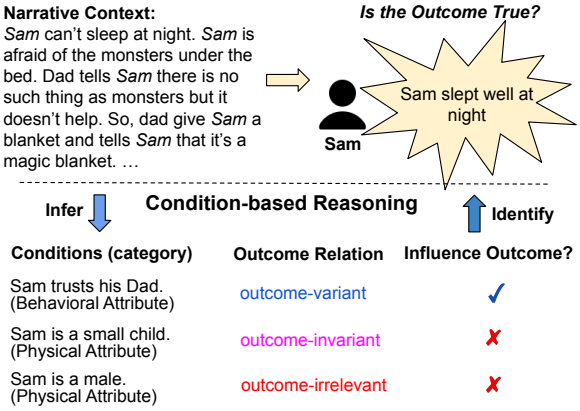
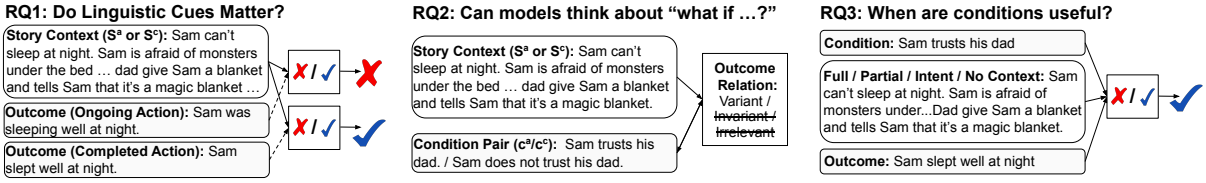


Figure 1: Conditions and their relationship to an outcome inspired by a story from PASTA/SAGA (Ghosh et al., 2023; Vallurupalli et al., 2024).

This type of reasoning is challenging because: (1) Conditions that to relate to entity properties and states are not always explicitly stated in a narrative but are implicitly understood through forming a coherent mental representation (Ghosh et al., 2023) and acquiring this implicit knowledge is not easy. (2) Outcomes can be vastly different even for slight variations in context (Vallurupalli et al., 2024) pointing to a need for identifying nuanced properties and states that have an influence on the outcome. (3) Knowing a condition's influence on the



(a) We examine whether different descriptions, an ongoing Vs a completed action, change the truth value of an outcome.

(b) We examine whether models are able to think counterfactually and generate/identify *outcome-variant* conditions.

(c) We examine whether models are able to combine a condition with varying amounts of context to validate outcomes.

Figure 2: Our research questions explore outcome validation through our condition-based reasoning tasks. We leverage PASTA (Ghosh et al., 2023) and SAGA (Vallurupalli et al., 2024) datasets to generate outcomes of SAGA and News Stories and validate these using both our generated and PASTA's crowd-annotated conditions.

outcome is implicitly understood through constructing counterfactual mental representations (Byrne, 2016; Mercier et al., 2017) requiring robust means for acquiring counterfactual conditions and reasoning with them. (4) Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022, *inter alia*), while powerful, do not necessarily perform well on tasks requiring counterfactual reasoning (Fang et al., 2025; Lin, 2004; Ghosh et al., 2023; Qin et al., 2019). (5) Linguistic differences influence action understanding (Pruś et al., 2024; Zhou et al., 2023; Salomon et al., 2013; Hart and Albarracín, 2011) which can hinder an LLM in understanding condition and outcome descriptions.

Inspired by Pruś et al. (2024)'s findings that pragmatic knowledge influences how linguistic cues are interpreted, we examine how models use linguistic cues when understanding story outcomes. We explore how implicit conditions influence outcomes and whether they can be used in place of missing context for validating story outcomes. We leverage two previously released datasets PASTA (Ghosh et al., 2023) and SAGA (Vallurupalli et al., 2024) consisting of participants' goal and state annotations (shown in Fig. 2) and augment them for exploring conditions and outcomes. We address the following research questions and formulate our **Condition-based Reasoning** tasks to explore the capabilities of intent-aligned LLMs on these tasks. We highlight the usefulness of conditions in determining story outcomes across a variety of alternative narratives, from short five sentence stories to real life newswire articles.

RQ1: Do Linguistic Cues Matter? We explore models' ability to handle different types of outcome descriptions. For the example in Fig. 1, would models consider "Sam was sleeping well" to mean the same as "Sam slept well?" While these two descriptions do not mean the same according to Dowty's imperfective paradox (Dowty, 1977), Pruś et al. (2024) show that meaning agreement is more complex. We examine how models fare on such linguistic differences and find that the state-of-the-art models differ in performance on the imperfective paradox, but understand both ongoing and completed actions, with a slight preference for completed actions. See §5.1 for more details.

RQ2: Can models think about "what-if...?" We examine whether models can reason counterfactually, through generating and identifying outcome-variant contrastive condition pairs. We found that with one exception (Mistral-7B-Instruct-v0.3), all considered models are better at identifying whether a condition is outcome-related (which does not involve counterfactual thinking), but struggle at identifying outcome variance which requires counterfactual thinking. See §5.2 for details.

RQ3: When are conditions useful? We examine if models are able to use conditions when context is unavailable or incomplete. We find that while models are able to combine conditions with available context to validate outcomes, some larger models are more cautious when the context is less constrained. See §5.3 for more details.

Overall, we find that performance on outcome validation improves by 2-6% (macro F1 score) when *outcome-variant* (as opposed to *outcome-invariant*) conditions are used with the story context. Since identifying or generating *outcome-variant* conditions is not easy, performance drops by 3-6% for generated conditions when compared to annotated conditions. For news stories, performance drops by up to 2-18%, when compared to the PASTA story generated conditions, likely attributable to a combination of the domain change and models' going beyond the provided context and using their knowledge of past news events from their pretraining. See §6 for more details.

Our contributions in this work include introducing the concept of conditions and their relationships

to an outcome. We artfully combine and bridge two existing datasets to explore intertwined aspects of conditions and outcomes. Through our research questions we examine how LLMs leverage conditions for outcome validation and compare their performance on the practical task of validating outcomes of News stories. Our data and code are available at https://github.com/saiumbc/CoRE.

2 Related Work

Examining the influence of states and actions on goal outcomes is an active research area in the cross-disciplinary fields of cognitive science and psychology (Hommel et al., 2001; Custers, 2023; Niv, 2019; Amir et al., 2024). The utility of counterfactual conditionals in reasoning (Goodman, 1947; Filho, 2012) has a long history in linguistics & psychology (Byrne and Tasso, 1999; Byrne, 2019).

States, Actions & Counterfactuals: The hypotheses in abductive natural language inference (α-NLI) task (Bhagavatula et al., 2020) and defeasible inference (δ-NLI) (Rudinger et al., 2020) are similar to the conditions we examine, however, these do not reference any specific goal or outcome. δ-NLI WIQA (Tandon et al., 2019) examined counterfactual situations through preturbing states with "what if" type of questions and obtained changes in action outcomes. PASTA (Ghosh et al., 2023) examined implied states and preturbed these states to examine changes in situational narratives. We condition PASTA states on goal achievement and obtain state conditions that influence goal outcomes.

Goals & Outcomes: Goal oriented reasoning has been explored in participant narratives (Rahimtoroghi et al., 2017), news actions (Jiang and Riloff, 2018) and procedural text (Zhang et al., 2020). LLM's goal reasoning has been explored using participant goals by Bellos et al. (2024) and SAGA (Vallurupalli et al., 2024). We extend SAGA to examine conditions' influence on goal outcomes.

Linguistic Understanding: Studying aspect for temporal reasoning has a long history in computational linguistics(Moens and Steedman, 1988; Siegel and McKeown, 2000). More recent work focused on aspect to study verb classes in text and their affect on textual entailment (Kober et al., 2020) and cross-domain data utility(Alikhani and Stone, 2019). Inspired by (Friedrich et al., 2023) and (Pruś et al., 2024) we examine LLMs' aspectual understanding for identifying goal outcomes.

3 Data Annotations

In the following paragraphs, we briefly describe the PASTA (Ghosh et al., 2023) and SAGA (Vallurupalli et al., 2024) datasets and how these datasets' annotations relate to conditions and outcomes.

(Counterfactual) State Condition of Participants in Narratives: The PASTA dataset is a collection of stories and implied states supported by the stories. For a given 5-sentence ROC Story (Mostafazadeh et al., 2016), crowd workers infer and describe a state implied by one or more of the sentences in the story (the annotations identify these sentences). For this state they also describe a perturbed state (a "what-if" type of counterfactual state) that is unsupported by the original story; they also minimally alter the story to support the perturbed state ensuring it does not support the original state. The dataset contains 5028 original stories and up to 3 state annotations per story, leading to 5715 state pairs and 5715 alternate stories. See Table 8 for examples of data annotations.

Goal Outcomes of Participants in Narratives: The SAGA dataset is a collection of goal related annotations for a subset of the PASTA story collection. Crowd workers describe an overarching goal of a volitional participant in the original story, how each actual story sentence relates to the goal and whether the goal is achieved; the described goal is what the participant hopes to achieve through their actions in the story and beyond. Alternate stories corresponding to the original story are examined to assess whether the actions in these stories can achieve the goal. The dataset contains three goal annotations for each participant in a story, for up to 4 story participants obtaining a total of 2785 goal annotations for 886 original stories (449 of these have 951 corresponding alternate stories). See Table 9 for examples of data annotations. We describe only the annotations we use in this work. For the complete list, please refer to Vallurupalli et al. (2024).

Conditions and Outcomes in Alternative Narratives: SAGA goals allow the examination of an entity's goal achievement in alternate stories and PASTA states[1] describe various properties and states of an entity that are supported by at least one of the alternate stories. Studying the interrelationships between these, we find that: (a) several different types of conditions can be inferred from

[1]We refer to PASTA states as conditions in this work.

a narrative, (b) not all conditions are related to a goal (these are *outcome-irrelevant*), and (c) slight changes in a narrative can lead to a contrastive condition but not all contrastive condition pairs lead to a different goal outcome. In some pairs, one state leads to achievement and the other not (we consider these to be *outcome-variant*) and in some pairs, both states lead either to goal achievement or not (we consider these to be *outcome-invariant*). See §4 for a detailed description with examples.

With this knowledge we augment PASTA and SAGA annotations for exploring conditions and outcomes. We consider an outcome to be a statement indicating the achievement (or not) of a participant's goal and want to identify whether the outcome is true for the story. For example in Fig. 1, the outcome is a statement about Sam achieving the goal of "sleeping well at night" and it is true for the story. We use a combination of prompting GPT3.5-Turbo and automatic methods to derive all annotations we need from PASTA/SAGA except for the outcome label. Our evaluation of conditions is based on outcome labels and annotating these requires understanding the story and deep reasoning. Hence we use an expert to obtain these label annotations. We describe our annotation process in Appendix A.5 listing the annotation statistics in Table 7 and data examples in Table 10.

4 Conditions & Outcomes

4.1 Conditions

Conditions inferred from a story include properties of entities ranging from the intrinsic (inherent attributes such as *Rob was immature*, *Cindy had long hair*, etc.) to the extrinsic (attributes resulting from other entities such as *Timothy has an old TV*, *Janice's closet is messy*, etc.). We expand the 3 categories used in PASTA (Ghosh et al., 2023), for error analysis on 200 random states on the story state inference task, to 4 categories and group all conditions into these as follows:

(a) **Physical:** This category includes natural physical attributes of an entity such as size, age, place etc., such as *Cindy had long hair*, *The cake is really big*, *The sky was cloudy etc*.

(b) **Functional:** This category includes attributes that influence an entity's capability to perform actions, e.g., *The machine was jammed*, *Timothy has an old TV*, etc.

(c) **Knowledge:** This category includes mental states of knowing information (knowledge about self or other entities, or pragmatic world knowledge), e.g., *Lisa knows her family well*, *Nancy is up to date with technology*, *Bill was aware that the cap had been loosened*, etc.

(d) **Behavioral:** This category includes behavioral aspects of an entity for example: *Cindy's dog is a biter*, *Charlie loves candies*, *the librarian is responsible*, *Amber is frugal*, etc.

This grouping identifies categories that are likely to be outcome-variant and useful for reasoning.

4.2 Outcome Relationships

The relationship between a condition and an outcome can be understood through counterfactual reasoning. Consider a pair of alternate story[2] contexts S_a and S_c that support a contrastive condition pair c_a (an actual condition) and c_c (a counterfactual condition) respectively where both stories contain goal-oriented events to achieve a common goal. As an outcome description O can be either true or false for a given story, we define 3 types of relationships:

A condition is **outcome-variant**, when O for S_a and S_c are different with one being true and the other being false. For the example in Fig. 1, we consider *Sam trusts his Dad* as c_a and *Sam does not trust his dad* as c_c. The outcome *Sam slept well at night* is true for S_a but false for S_c. Hence, both conditions are outcome-variant.

A condition is **outcome-invariant**, when outcome O is true (or false) for both S_a and S_c. For the example in Fig. 1, *Sam is a small kid* is the c_a and *Sam is a big kid* the c_c. Both conditions are outcome-invariant because the outcome *Sam slept well at night* is true for both contexts, S_a and S_c. While these conditions can be seen as outcome-variant through the use of commonsense inference such as *a small kid trusts their parent*, in such cases, the commonsense inferred condition will be the outcome-variant condition.

A condition is **outcome-irrelevant** if it and its counterfactual condition have no bearing on the outcome. For the example in Fig. 1, the actual condition *Sam is a boy* and the corresponding counterfactual condition *Sam is a girl* have no relevance to the outcome *Sam slept well at night*.

[2]We use "narrative" and "story" interchangeably, even though we acknowledge they have important differences.

4.3 Lexical Cues & Imperfective Paradox

Our outcomes describe the end-result of volitional actions. In Fig. 1 Sam slept well is the outcome of several goal-oriented actions. This outcome can be interpreted as completed or ongoing (incomplete) based on linguistic cues such as verb aspect and world knowledge (Givón, 1992; Magliano and Schleich, 2000; Madden and Zwaan, 2003) and annotated as such. Whether an outcome is true depends on accurately deciphering whether the planned goal is achieved, regardless of linguistic differences. In Fig. 1, whether the outcome is described as Sam slept well or Sam was sleeping well, we want to know whether Sam achieved the goal "of sleeping at night."

Imperfective Paradox: Dowty (1977) notes that according to the "Imperfective Paradox," an accomplishment described as a past ongoing action is not necessarily the same as a past completed action, i.e., that Sam slept well is not the same as Sam was sleeping well (we consider *sleeping well* and *slept well* in the context of Sam's goal accomplishment).

Usefulness of Aspect: The semantic property of lexical aspect associated with verbs helps us understand how actions unravel over time (Smith, 1983; Vendler, 1957). Grammatical aspect helps us distinguish between a completed and an ongoing action through analyzing different view-points–of the entire situation vs. a part of it (Smith, 1999). Lexical aspect helps us understand *sleep* as an activity and grammatical aspect helps us consider when *slept* and *sleeping* mean the same. (See Friedrich et al. (2023) for a more detailed discussion).

Our study: Pruś et al. (2024) show that pragmatic reasoning rooted in world knowledge influences how aspect is understood and that an LLM's pragmatic world knowledge (acquired during pretraining) leads to a different aspectual understanding than that of humans. We extend their study to examine to what extent LLMs' aspectual knowledge affects condition and outcome understanding, using the imperfective Paradox with and without story context. Specifically, we examine if LLMs consider 'Sam was sleeping well' to be different from 'Sam slept well' for determining if Sam's goal is achieved for several alternate story contexts and when no context is provided. We examine whether models can leverage the provided context and perform a pragmatic view-point analysis that is dictated by the context to improve their inferences.

5 Outcome Validation through RQs

We examine whether an outcome is possible for a condition known about an entity through our research questions. We use both crowd-annotated and LLM-generated conditions, where not every condition is supported by the story context or related to the outcome. We formulate reasoning tasks with the aim to compare various models' performance on identifying and generating outcome-variant conditions. We examine the conditions' usefulness in validating outcomes of SAGA and News Stories.

Models: We examine well-known closed and open, human intent-aligned LLMs of different sizes: GPT-4o-mini, GPT-4o, FlanT5-XXL, LLaMA-3.1-8B-Instruct (Llama-8BI), Mistral-7B-Instruct-v0.3 (Mistral-7BI) and LLaMA-3.1-70B-Instruct (Llama-70BI). We refer to the models by the shortened names in parentheses. We examine the impact of model type, size and the alignment.

Inference: For the tasks examined in RQ1 and RQ3, we use zero-shot prompting to examine models' inherent knowledge. For the tasks examined in RQ2 we use few-shot prompting which is required as models do not inherently understand how to generate or identify outcome-variance. Inference cost for GPT models was less than $100.

5.1 RQ1: Do Linguistic Cues Matter?

We examine aspect understanding using two sets of outcome descriptions (past ongoing & past completed actions) using (a) Dowty's imperfective paradox (discussed in §4.3) and (b) direct questioning. See prompts for both tasks in Table 12.

We compare performance on the imperfective paradox with and without the story context to examine whether models are able to improve upon their aspectual understanding leveraging the provided context. We extend the prompt from Pruś et al. (2024) to (i) include an *Unsure* option, in addition to *Yes* and *No*, which is useful when stories do not provide enough information relating to an outcome, and (ii) optionally provide story context.

With direct questioning, we examine if models have a preference for a specific type of linguistic cue and if their understanding of the two descriptions agrees with that of human preference. For the latter, we compute Cohen's Kappa (κ) between the labels for both descriptions.

Evaluation: We compare the F1 values for the 3 labels (True/False/Unsure) of the imperfective

Model	No Context	Story Context
Flan-T5-XXL	.68/.00/.09	.71/.21/.00
GPT-4o-mini	.64/.27/.00	.73/.49/.02
GPT-4o	.37/.50/.15	.69/.64/.11
Mistral-7BI	.69/.08/.00	.47/.61/.05
Llama-8BI	.48/.44/.00	.23/.53/.00
Llama-70BI	.41/.46/.00	.44/.56/.00

Table 1: F1 scores for True/False/Unsure answers to the Dowty's Imperfective Paradox with ('Story Context' column) and without ('No Context' column) the story context. Access to context improves Flan-T5 & GPT.

paradox without and with story context. We derive the gold labels for this task as follows: we use the gold label from the on-going action, when the label for the completed action is true, otherwise, we use the label for the completed action. For direct questioning, we compare F1 values for all 3 labels prompting twice for the two outcome descriptions. We compute Cohen's Kappa (κ) between model generated labels and gold labels for both prompts.

Results & Discussion: According to results from the imperfective paradox (see Table 1), FlanT5-XXL, GPT-4o-mini and Mistral-7BI are better than GPT-4o and Llama models at aspectual understanding. FlanT5-XXL, GPT-4o-mini and GPT-4o are able leverage context to improve upon their base aspectual understanding with GPT-4o improving by a large amount; Mistral 7BI and the Llama models are unable and perform poorly. This performance indicates how models handle linguistic complexities in our tasks (see examples in Table 13).

The results from direct questioning (in Table 2) show that all models except GPT-4o have a slight preference towards completed action descriptions with their better performance. This is true for Llama and Mistral models as well despite their poor performance on the imperfective paradox shown in Table 1. The high agreement between in-progress and completed descriptions, for all models except GPT-4o, is comparable to the agreement seen in gold labels (within $\pm$ .06). Some models like FlanT5-XXL do not distinguish the linguistic nuance between the two descriptions and others like the Llama and GPT models distinguish more. GPT-4o is more cautious and generates an 'Unsure' label in nuanced situations. See examples in Table 14.

5.2 RQ2: Can Models Think "What If ...?"

We examine whether models can think counterfactually, when generating and identifying conditions,

Model	In-progress Action	Completed Action	Cohen's kappa κ
Gold Label	-	-	.77
FlanT5-XXL	.81/.67/.12	.82/.71/.19	.84
GPT-4o-mini	.80/.72/.11	.81/.77/.24	.74
GPT-4o	.80/.79/.46	.80/.79/.46	.63
Mistral-7BI	.75/.68/.02	.76/.70/.10	.77
Llama-8BI	.77/.68/.00	.79/.72/.00	.71
Llama-70BI	.75/.69/.05	.80/.74/.19	.72

Table 2: F1 values for models output of True/False/ Unsure with directly questioning if an outcome description is true for a story context and the two sets of outcome descriptions. The κ column shows agreement between labels generated for both descriptions (and gold labels).

formulating both generation and classification tasks as follows: (a) We examine if models can generate contrastive condition pairs that are *outcome-variant* (when prompted with similar examples) and *outcome-invariant* (again when prompted with similar examples) (b) We examine whether models can identify a condition's outcome-relevance and a contrastive pair's outcome variance.

5.2.1 Generating Outcome Relationships

We prompt models with a story and an outcome and generate contrastive condition pairs where the first condition is expected to be supported by the story and the second condition is expected to be a counterfactual condition (see Table 15 for the prompt). To ensure, the conditions are contrastive, we use incontext learning examples.

Evaluation: We manually examine 100 random condition pairs selecting an equal number of original and alternate stories and an equal number of pairs expected to be outcome variant and invariant. Our evaluation rubric is designed to obtain a binary answer of Yes or No and consists of the following questions. (i) *Con.*: Are conditions in the pair contrastive? (ii) *Rel.*: Is least one of the conditions outcome relevant? (iii) *Sup.*: Is the first condition in the pair supported by the story? (iv) *Var.*: Is the pair outcome-variant for the story context? When a single condition in the pair is outcome relevant, it should be supported for the pair to be relevant.

Results & Discussion: We show an example prompt and the conditions generated by the various models in Table 3. Generating contrastive and supported conditions that are outcome- relevant and variant indicate a model's counterfactual thinking and have an impact on outcome validation (See Table 16 for erroneous generations). Using

Prompt for generating a condition pair	Model & Generated condition pair
A: Sam can't sleep at night. He is afraid of the monsters under the bed. Dad tells him there is no such thing as monsters but it doesn't help. So Dad gives him a blanket and tells him that it's a magic blanket. Sam believes that he is protected by the blanket and he sleeps well. **B:** Sam had a restful night's sleep. Generate a pair of contrastive conditions relevant to the context in A and the statement in B. Please make sure the conditions are stative expressions and do not describe activities.	answer: Sam didn't believe that the blanket was a magic blanket. [FlanT5-XXL] Sam feels anxious about sleeping., Sam feels calm about sleeping. [GPT-4o-mini] Sam's belief in the magic of the blanket is strong.,Sam's belief in the magic of the blanket is weak. [GPT-4o] Sam is afraid of monsters under the bed., Sam had a peaceful sleep. [Mistral-7BI] Sam feels protected from the monsters, Sam feels vulnerable to the monsters [Llama-8BI] Sam feels unsafe at night, Sam feels safe at night [Llama-70BI]

Table 3: Examples of Model Generated Conditions. Unlike other models which generate fairly good condition pairs, FlanT5-XXL does not generate a pair and Mistral-7BI does not generate a contrastive and variant pair.

Model	Con.	Rel.	Sup.	Var.
FlanT5-XXL	10	88	66	6
GPT-4o-mini	96	100	80	100
GPT-4o	91	97	82	94
Mistral-7BI	94	100	94	88
Llama-8BI	94	100	86	94
Llama-70BI	100	100	96	100

Table 4: Evaluation of generated condition pairs on 4 factors using 100 random generations from each model. Llama-70BI generates condition pairs of high quality.

exclusively *outcome-variant* or *outcome-invariant* did not ensure the generation of condition pairs of the same type. Our manual evaluation results (in Table 4) show that models are unable to distinguish between the two types and always generate a mix of mostly outcome-variant conditions. The pairs are mostly contrastive ($< 10\%$ are not) and outcome-relevant but are not always ordered (the first condition being the one supported by the story and the second being a counterfactual). All models, except, FlanT5-XXL and Llama70bI, are better at generating a pair that is ordered or outcome-variant but Llama70BI is better at both. FlanT5-XXL is poor in all evaluated criteria, however, the generated condition contains information highly useful for outcome validation (see §5.3 results).

5.2.2 Identifying Outcome Relationships

To identify a condition pair's outcome variance, we examine both a single-step reasoning and a multi-step reasoning where we using a standard Chain-of-Thought (CoT) (Wei et al., 2022) approach (use gold labels for each condition's relevance and the pair's variance to reason step by step in the in-context example with the model performing the same during inference) and a variant approach where each condition's outcome relevance from two additional prompts is provided at inference time. See Table 15 for these prompts.

Model	Single-step	Stand. CoT	Alter. CoT
FlanT5-XXL	.43/.55	.46/.56	.38/.62
GPT-4o-mini	.54/.38	.50/.47	.42/.59
GPT-4o	.52/.59	.52/.62	.41/.71
Mistral-7BI	.32/.67	.00/.72	.09/.72
Llama-8BI	.53/.35	.38/.60	.55/.24
Llama-70BI	.57/.37	.54/.36	.54/.48

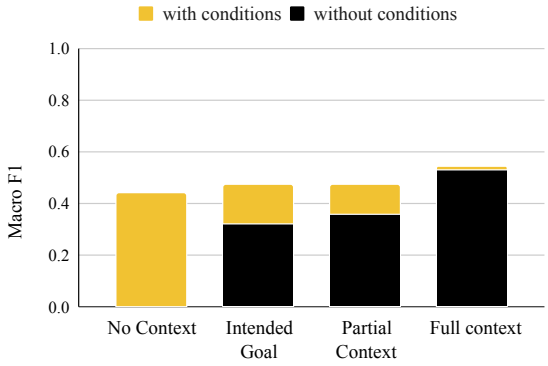
Table 5: Identifying a condition pair's outcome-variance is difficult for models. We compare F1 scores for whether a condition pair is outcome- variant/invariant using single and multi-step reasoning. Standard CoT uses gold labels for in the in-context examples; alternative CoT uses answers from condition-relevance (see Table 6) during inference.

Model	Irrelevant	Relevant	Support	Oppose
FlanT5-XXL	.02	.87	.55	.01
GPT-4o-mini	.02	.86	.44	.57
GPT-4o	.28	.78	.32	.54
Mistral-7BI	.37	.04	.05	.00
Llama-8BI	.00	.87	.42	.55
Llama-70BI	.00	.87	.48	.59

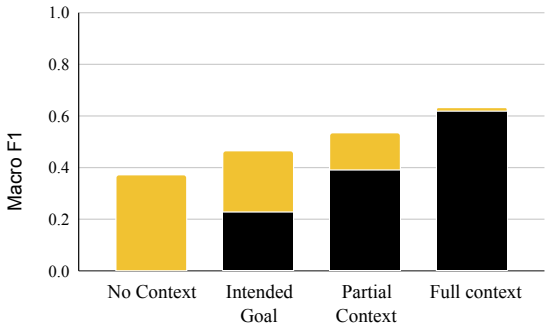
Table 6: Identifying a condition's relevance to the outcome is easier for models but they struggle to identify whether a condition supports or opposes the outcome. Identifying whether a pair is variant depends on the support/oppose relation from the intermediate step of Alternate CoT lowering its performance.

Evaluation: We compare performance of the single-step and multi-step approaches using the F1 values for each task label. We interpret the 3 labels of True/False/Unsure for binary outcome-relevance and variance as follows: True and False labels are relevant (i.e., supporting & opposing relations) and unsure labels are irrelevant. For variance, True labels are variant, False & Unsure labels are invariant. We consider these groupings appropriate for positively identifying variance and relevance.

Results & Discussion: Models find it difficult to identify outcome- variant and invariant conditions, as seen by the lower F1 scores for both labels (see Table 5). GPT-4o and Llama-70BI are slightly bet-



(a) **Llama-8BI** leverages conditions to improve perfor-



(b) **Llama-70BI** is cautious when using partial contexts.

Figure 3: Models are able to utilize conditions in addition to the story context to improve performance when identifying outcome labels. The black bars show performance with story context alone and the yellow bars show the improvement with annotated variant conditions used in conjunction with the story context. We show Llama models here and the other models in Fig. 7.

ter than the smaller models with Mistral-7BI being the worst. The standard CoT prompting improves only FlanT5-XXL with other models only improving in identifying invariance. Performance for the alternative CoT prompt suffers from relying on an intermediate step to identify a condition's support. Our results (in Table 6) show that all models except Mistral-7BI are good at identifying whether a condition is relevant but are poor at identifying whether it supports or opposes the outcome. This is likely due to the nature of our alternate stories where minimal changes can alter the outcome.

5.3 RQ3: When are conditions useful?

We examine whether models are able to use conditions when available story context is incomplete. We identify if the outcome is true when given varying amounts of story context and a condition (C for crowd-sourced or GC for generated) as follows:

1. full story context (Full SC+C, Full SC+GC),
2. partial story context after removing the sentences from which the condition was derived (Par SC+C, Par SC+GC),
3. indicating the participant's intent of achieving their goal (Goal Int+C, Goal Int+GC),
4. no story context (C, GC).

We use the baselines of only the story context (Full SC), the partial story context (Par SC) and the intended goal (Goal Int). See prompts in Table 17.

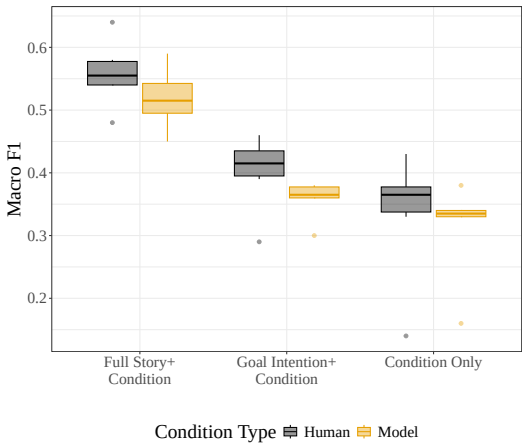


Figure 4: **Model** generated conditions are not as good as **Human** annotated conditions leading to lower performance on outcome identification. We compare task settings with varying **SAGA story** contexts paired with an annotated or a generated condition. Results are aggregated across all models.

Evaluation: We compare the above settings using Macro F1 for True/False/Unsure labels.

Results & Discussion: We find that models are able to leverage the condition as shown in Fig. 3 and Fig. 7. Performance improves when outcome-variant conditions are paired with the context. Some models are able to leverage additional information provided by implied conditions even when the full story context is available. Performance across all settings is better for outcome-variant conditions (see Table 18). When the context is unconstrained, GPT-4o, Llama-70BI and Mistral-BI are more cautious (selecting 'Unsure' instead of 'Yes' or 'No' for the incomplete contexts). FlanT5-XXL's higher performance is attributable to StoryCloze (similar to our outcome identification task) being a part of Flan's 1.8K fine-tuning tasks.

Influence of Generated Conditions We compare performance when using model generated conditions vs. all annotated conditions. Performance is reflective of a model's condition generation capability discussed in §5.2.1. As seen in Fig. 4, generated conditions lead to .03 - .05 drop in Macro F1 (averaged across models) for the different context settings. See Table 19 for all models' performance.

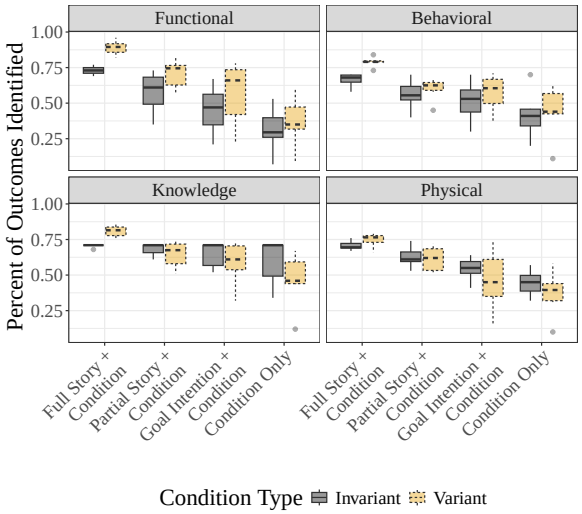


Figure 5: Models are better at utilizing functional conditions better than physical conditions for reasoning about outcome labels. We identify outcome labels using various story contexts and a variant or an invariant condition pair where we track the category of the first condition. Each boxplot aggregates over all six models.

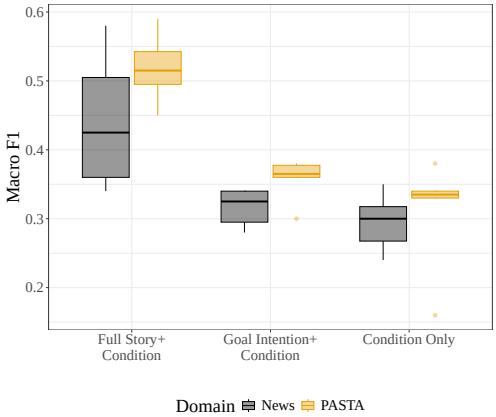


Figure 6: We compare the performance of outcome identification for **PASTA/SAGA stories** and **SAGA News stories** using varying story contexts and generated conditions for both story types. News stories are more complex leading to a drop in performance when using only the story context. Conditions used with the story context fill some of the performance gap. Results are aggregated across all models.

Influence of Condition Category We examine which categories of conditions help models reason in both complete and incomplete story contexts. As seen in Fig. 5 performance across all groups is better for outcome- variant than invariant conditions. For outcome-variant conditions, functional, knowledge and behavioral conditions lead to better performance than the physical when identifying outcome labels. The first condition in a condition pair determines the category and the pair determine the outcome relation.

Error Analysis: When a condition pair is not contrastive, or contrastive but not outcome variant, it can lead to poorer performance on outcome identification. This is worse in partial context settings where a model fails to fill in the missing context because it is unable to align the conditions with the context. We show examples of such misaligned conditions in Table 20. Performance drop in incomplete context settings is mostly seen with Functional and Behavioral conditions than knowledge and physical categories (shown in Fig. 5). We normalize outcome identification errors within each condition category.

6 Cross-Domain Application to News

We use the task prompts and setup from RQ2 and RQ3 to generate conditions for News stories and use them to identify outcome labels. News stories have longer sentences and include more complexities in the number of entities, topics, linguistic cues etc. when compared to SAGA stories. We examine their affect on model performance, comparing generated conditions from both story types. We compare the settings in RQ3, except for partial story context, as this requires identifying story sentence(s) supporting the condition.

Results & Discussion: The complexity of news increases the difficulty of the task and bias from models' knowledge of past events leads to a performance drop of .11 when using the full story context (see Table 22). When using conditions with the context, the drop is .02 -.08 as shown in Fig. 6. See examples of model generations in Table 21.

7 Conclusions

By defining *Condition-based Reasoning tasks*, we have demonstrated the importance of identifying outcome-oriented relationships and their usefulness in validating story outcomes. To do this, we first combined and augmented two existing datasets to obtain outcome- variant and invariant conditions, and then showed that models can effectively use outcome-variant conditions in place of missing context. We also showed how to generate and use outcome-variant conditions on a new domain. We examined and showed models' sensitivity to linguistic cues, which can indirectly affect both condition generation and outcome validation. We hope that our work will help further linguistic examination of model behaviors to achieve a deeper and more robust narrative understanding.

8 Limitations

We acknowledge our work has the following limitations:

1. The linguistic analysis we performed is limited to the verbs used in our outcome descriptions.

2. The pre-trained large language models in our experiments can echo biases and misinformation either implicitly or explicitly. We do not attempt to control for these in this work.

3. We focus on more formal written english and our annotations are based on well known NLP data sources. We use pre-trained large languages models to rewrite, generate and evaluate these annotations for use in our experiments.

4. While we manually evaluated a subset of the generated data and did not find any misinformation, it is possible for the generated data to contain misinformation.

5. Our annotation and reasoning tasks do not examine biases in the data sources and hence do not control for them.

6. Model generation and classification abilities can vary with the formality, style, and mood in the crowd written stories we annotated.

Acknowledgments

We wish to thank the anonymous reviewers for their helpful comments, feedback, and suggestions. We would also like to thank Katrin Erk, Sayontan Ghosh and Niranjan Balasubramian for early discussions and feedback. 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 DARPA for the SciFy program under agreement number HR00112520301. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any 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 ARL, DARPA or the U.S. Government.

References

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadav Amir, Yael Niv, and Angela J Langdon. 2024. States as goal-directed concepts: an epistemic approach to state-representation learning. In *Reinforcement Learning Conference*.

Filippos Bellos, Yayuan Li, Wuao Liu, and Jason Corso. 2024. Can large language models reason about goal-oriented tasks? In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 24–34, St. Julian's, Malta. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020 - Addis Ababa, Ethiopia*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, 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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Ruth M.J Byrne. 2016. Counterfactual thought. *Annual Review of Psychology*, 67:135–157.

Ruth MJ Byrne. 2019. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pages 6276–6282. California, CA.

Ruth MJ Byrne and Alessandra Tasso. 1999. Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & cognition*, 27:726–740.

Gergely Csibra and György Gergely. 2007. 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, 124(1):60–78. Becoming an Intentional

Agent: Early Development of Action Interpretation and Action Control.

Ruud Custers. 2023. Thoughts about actions and outcomes (and what they lead to). *Motivation Science*, 9.

David R. Dowty. 1977. Toward a semantic analysis of verb aspect and the english 'imperfective' progressive. *Linguistics and Philosophy*, 1(1):45–77.

Yi Fang, Moxin Li, Wenjie Wang, Lin Hui, and Fuli Feng. 2025. Counterfactual debating with preset stances for hallucination elimination of LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10554–10568, Abu Dhabi, UAE. Association for Computational Linguistics.

Oswaldo Filho. 2012. Goodman and parry on counterfactual. *Principia: an international journal of epistemology*, 15.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2023. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 599–622, Dubrovnik, Croatia. Association for Computational Linguistics.

Sayontan Ghosh, Mahnaz Koupaee, Isabella Chen, Francis Ferraro, Nathanael Chambers, and Niranjan Balasubramanian. 2023. PASTA: A dataset for modeling PArticipant STAtes in narratives. *Transactions of the Association for Computational Linguistics*, 11:1283–1300.

T. Givón. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics*, 30(1):5–56.

Nelson Goodman. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy*, 44(5):113–128.

William Hart and Dolores Albarracín. 2011. Learning about what others were doing: Verb aspect and attributions of mundane and criminal intent for past actions. *Psychological Science*, 22(2):261–266.

Bernhard Hommel, Jochen Müsseler, Gisa Aschersleben, and Wolfgang Prinz. 2001. The theory of event coding (tec): A framework for perception and action planning. *The Behavioral and brain sciences*, 24:849–78; discussion 878.

Tianyu Jiang and Ellen Riloff. 2018. Learning prototypical goal activities for locations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1297–1307, Melbourne, Australia. Association for Computational Linguistics.

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality across genre: A distributional semantics approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on 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.

Carol J Madden and Rolf A Zwaan. 2003. How does verb aspect constrain event representations? *Memory & cognition*, 31(5):663–672.

Joseph P Magliano and Michelle C Schleich. 2000. Verb aspect and situation models. *Discourse processes*, 29(2):83–112.

Hugo Mercier, Jonathan J. Rolison, Marta Stragà, Donatella Ferrante, Clare R. Walsh, and Vittorio Girotto. 2017. Questioning the preparatory function of counterfactual thinking. *Memory and Cognition*.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen. 2016. Story cloze evaluator: Vector space representation evaluation by predicting what happens next. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 24–29, Berlin, Germany. Association for Computational Linguistics.

Yael Niv. 2019. Learning task-state representations. In *Nature neuroscience*, pages 1544–1553.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Katarzyna Pruś, Mark Steedman, and Adam Lopez. 2024. Human temporal inferences go beyond aspectual class. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1923, St. Julian's, Malta. Association for Computational Linguistics.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. 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 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369, Saarbrücken, Germany. Association for Computational Linguistics.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Meghan M Salomon, Joseph P Magliano, and Gabriel A Radvansky. 2013. Verb aspect and problem solving. *Cognition*, 128(2):134–139.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning methods to combine linguistic indicators:improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–627.

Carlota S. Smith. 1983. A theory of aspectual choice. *Language*, 59(3):479–501.

Carlota S. Smith. 1999. Activities: States or events? *Linguistics and Philosophy*, 22(5):479–508.

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.

Sai Vallurupalli, Katrin Erk, and Francis Ferraro. 2024. SAGA: A participant-specific examination of story alternatives and goal applicability for a deeper understanding of complex events. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15396–15420, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Appendix

A.1 AI Assistance

We did not use any AI assistants for writing the paper or for any of the coding used in our experiments. All writing is original and entirely produced by the authors.

A.2 Infrastructure

We used both RTX 8000 with 48GB of GPU memory and Nvidia A100 with 80GB GPU memory for inference with and without in-context examples. Run time for a model ranges from a .52 hours.

A.3 Hyperparameters

We use a temperature of .6 for non-GPT models (1 for GPT models) and nucleus sampling to get the top 10%. With these settings we were able to generate a variety of conditions that are closely related to the story. We obtain multiple sufficiently diverse conditions through repeated sampling using these settings. For simplicity, we use the same settings for all the tasks.

For in-context learning examples, we use a label balanced set of 2 examples for each of the Yes/No/Unsure when generating or identifying variant conditions. Additionally, we only use alternate stories and their implicit conditions to ensure examples were unseen by the foundation models.

A.4 PASTA & SAGA Data

We provide examples of PASTA & SAGA annotations we utilize in Table 8 and Table 9. We also use an SAGA News dataset of 250 goal annotations.

A.5 Data Augmentation

We made the following augmentations to SAGA & PASTA annotations for conditions and outcomes. We provide annotation statistics in Table 7 and annotation examples in Table 10 and Table 11.

(1) We rewrote the SAGA goal descriptions as ongoing and completed actions using GPT-3.5-Turbo. SAGA goals are not always grammatically correct and complete sentences; we wrote 3 seed examples (without using any story context) and with these rewrote 100 SAGA goals from the training data split as complete grammatical sentences describing past completed and ongoing actions. The story context is not used for this rewrite, to ensure it is only a sentence rewrite without any story based reasoning. The authors manually verified that the generated claims are grammatically correct sentences reflecting a completed goal relevant to the story. Using the verified examples we rewrote the remaining goals and manually verified using the same checks as with the first 100.

(2) Asking whether the outcomes (both ongoing and completed actions) from the above step can be inferred from the story context (using both actual and alternate story contexts of SAGA), we manually annotated whether each outcome is true, false or unsure for the given story. We computed agreement with the crowd annotations in step (4) to show that these manual annotations are in fairly good agreement with the crowd annotations.

(3) The goal achievement annotation in SAGA based on 5-point likert scale from Unsuccessful to Fully Successful is not always accurate. We used the goal achievement evaluation scores from the 3 crowd workers for the test and validation data splits to automatically correct the achievement annotation

(4) We convert the 5-point likert scale to the 3 outcome labels of True/False/Unsure (Fully & Moderately Successful labels are converted to an outcome label of true and Less & Un-Successful labels are converted to an outcome label of false). We compute IAA between our outcome label annotations and these automatically corrected crowd annotations and achieve a Cohen's Kappa score of .64 (without the corrections this score is .54). A majority of disagreement is due to unsure labels and if we were to consider only the true and false labels the IAA is .73 (without the corrections

Annotation	Augmented SAGA/PASTA
Story Contexts	886 (Actual) / 951 (Alternate)
Claims	2985
# Condition Pairs	951
# Condition applied to Claim	6989
Claim Labels (identify claim truthfulness)	2125 / 299 / 561 (True/False/Unknown)
Corrected Goal Labels (identify goal achievement)	2125 / 484 / 376 (Success/Unsuccess/Unsure)
Event-Goal Relations (5 relations per claim)	13140 / 1230 / 555 (Support/Oppose/No-Effect)
Condition Type (condition-outcome relation)	6607 / 3448 / 3923 (Support/Oppose/No-Effect)
Outcome-oriented Relations (a condition-pair's relation)	1985 / 694 / 598 (variant/invariant/unsure)

Table 7: Augmented SAGA & PASTA Data Statistics.

PASTA Annotation	Annotated Data
Original story 1 (story presented to annotators)	Susan wanted to start a business. She decided to start an Italian restaurant in her home town. She hired a world famous chef to lead her kitchen. Her restaurant eventually became very successful. It was featured in many magazines and TV shows.
Implied state #1:	Susan likes Italian food.
Supporting sentences:	2
Preturbed state #1:	Susan hates Italian food.
Supporting sentences:	2
Altered story supporting preturbed state #1 **(Alternate Story #1)**:	Susan wanted to start a business. She decided to start a Chinese restaurant in her home town. She hired a world famous chef to lead her kitchen. Her restaurant eventually became very successful. It was featured in many magazines and TV shows.
Implied state #2:	The world famous chef knows Italian food very well.
Supporting sentences:	2,4
Preturbed state #2:	The world famous chef does not know Italian food at all.
Supporting sentences:	4,5
Altered story supporting preturbed state #2 **(Alternate Story #2)**:	Susan wanted to start a business. She decided to start an Italian restaurant in her home town. She hired a world famous chef to lead her kitchen. Her restaurant was eventually a failure because the food was not Italian at all. She missed the chance to feature in many magazines and TV shows.
Original story 2 (story presented to annotators)	Eli predicted the stock market trends on a lark. When every prediction came true his friends were in awe. They asked him to do it for the next day. His predictions turned out to be sheer luck. Eli's friends were angry when they lost money on their purchases.
Implied state #1:	Eli does not know anything about the stock market.
Supporting sentences:	1,4
Preturbed state #1:	Eli is a highly talented stock analyst.
Supporting sentences:	1,4,5
Altered story supporting preturbed state #1 **(Alternate Story #1)**	Eli predicted the stock market trends as part of his job. When every prediction came true his friends were in awe. They asked him to do it for the next day. His predictions turned out to be ingenious. Eli's friends praised him when their purchases went up in value.

Table 8: Examples of **PASTA crowd annotations** showing two original stories. Each original story can have upto 3 pairs of Implied & Preturbed state pairs (these examples have 2 and 1 pairs respectively). Crowd workers infer an implied state and the story sentences supporting it. They describe a counterfactual preturbed state and minimally alter the story sentences to support the preturbed state (altered sentences support the preturbed state).

this score is .63) showing that our claim label annotations are in good agreement with the crowd, especially after applying corrections based on multiple worker evaluations.

(5) Each of the 5 sentences in a SAGA actual story was assigned one of 6 labels to identify the relation between the event in the sentence and the goal by a crowd worker. We map these 3 relation labels as follows to identify the relationship between the sentence and outcome: any goal justifying and enabling relations are labeled *Support*, any goal blocking relation is labeled *Oppose* and the remaining are labeled *No-Effect*. We realize this mapping can be noisy, especially for data in the training split which could not be corrected in the above step (the evaluation scores were only available for test and validation data splits). Using the SAGA sentence relations we automatically identify one of the 3 above relations for each condition (using the PASTA sentence annotations that identify which sentences a condition is based upon). We assume the counterfactual condition would have the opposite relation which can be noisy. These relations identify how a condition relates to the outcome (aka outcome-oriented relation annotations).

6. We automatically identify if a pair of contrastive conditions are outcome- variant, invariant or unsure using the goal achievement annotations from the two stories.

7. We use Spacy to obtain the root verb of PASTA states (aka conditions) and categorize these descriptions into stative and action-oriented expressions to identify the lexical type of the annotation. About a third are activity oriented expressions and two-thirds are stative expressions. We also categorize the conditions using the 4 categories from §4.

8. We obtain a random 125 news stories from an extended version of the previously published SAGA (Vallurupalli et al., 2024) annotation effort. These news stories were selected randomly from the National, Foreign and Financial news desks of the Annotated New York Times (ANYT) newswire dataset to obtain stories that contain several participants involved in situations reflecting the complexity of common everyday situations.

9. For the news stories, we followed the same

Annotation	Data
Original story 1 (presented to annotators)	Susan wanted to start a business. She decided to start an Italian restaurant in her home town. She hired a world famous chef to lead her kitchen. Her restaurant eventually became very successful. It was featured in many magazines and TV shows.
Volitional Participant: Susan's Goal: Goal Achievement in Original story: Alternate story #1: Alternate story #2: Story Sentence & relationship to Goal (obtained only for original story)	Susan Create a successful business Fully Successful Fully Successful Unsuccessful 1->Justifies some aspect of the goal 2->Enables or Helps to potentially achieve some aspect of the goal 3->Enables or Helps to potentially achieve some aspect of the goal 4->is an effect caused by an action related to the goal 5->Enables or Helps to potentially achieve some aspect of the goal
Original story 2 (presented to annotators)	Eli predicted the stock market trends on a lark. When every prediction came true his friends were in awe. They asked him to do it for the next day. His predictions turned out to be sheer luck. Eli's friends were angry when they lost money on their purchases.
Volitional Participant: Eli's Goal: Goal Achievement in Original story: Alternate story #1: Story Sentence & relationship to Goal (obtained only for original story)	Eli Advise his friends on their finances. Unsuccessful Fully Successful 1->is Related to another sentence, but, unrelated to the goal 2->Justifies some aspect of the goal 3->Justifies some aspect of the goal 4->Enables or Helps to potentially achieve some aspect of the goal 5->is an effect caused by an action related to the goal
Volitional Participant: Eli's friends's Goal: Goal Achievement in Original story: Alternate story #1: Story Sentence & relationship to Goal (obtained only for original story)	Eli's friends to make some quick easy for sure money. Unsure Fully Successful 1->is Related to another sentence, but, unrelated to the goal 2->Justifies some aspect of the goal 3->Enables or Helps to potentially achieve some aspect of the goal 4->Prevents or Blocks the achievement of some aspect of the goal 5->is an effect caused by an action related to the goal

Table 9: Examples of **SAGA crowd annotations** showing two original stories. Each original story can have goals annotated for upto 4 participants (these examples have 1 and 2 participants respectively). Crowd workers describe an overarching goal for a participant and identify whether the goal is achieved in the story and in all the alternate stories. They also identify how each story sentence relates to the goal.

process as in steps 1 and 2 above to obtain an outcome and its truth label.

10. We do not annotate conditions, but obtain conditions using the condition generation task in §5.2.1.

Annotation	Data
Original Story 1	
Outcome described as an ongoing activity:	Susan was creating a successful business.
Outcome Label (Original story):	True
Outcome described as a completed activity:	Susan created a successful business.
Outcome Label (Original story):	True
Outcome Label (Alternate story 1):	True
Outcome Label (Alternate story 2):	False
For condition pair #1	
Story Sentence relationship for Implied state #1:	Enables
Story Sentence relationship for Preturbed state #1:	Opposes
Outcome Variance:	Outcome-Invariant
For condition pair #2	
condition-outcome relationship for Implied state #2:	Enables
condition-outcome relationship for Preturbed state #2:	Opposes
Outcome Relationship	Outcome-Variant
Original Story 2	
Outcome described as an ongoing activity:	Eli was providing financial advice to his friends.
Outcome Label (Original story):	True
Outcome described as a completed activity:	Eli provided financial advice to his friends.
Outcome Label (Original story):	True
Outcome Label (Alternate story 1):	True
For condition pair #1	
condition-outcome relationship for Implied state #1:	Enables
condition-outcome relationship for Preturbed state #1:	Opposes
Outcome Relationship:	Outcome-Invariant
Outcome described as an ongoing activity:	Eli's friends were finding a quick and easy way to make some money.
Outcome Label (Original story):	False
Outcome described as a completed activity:	Eli's friends found a quick and easy way to make some money.
Outcome Label (Original story):	False
Outcome Label (Alternate story 1):	True
For condition pair #1	
condition-outcome relationship for Implied state #1:	Opposes
condition-outcome relationship for Preturbed state #1:	Enables
Outcome Relationship:	Outcome-Variant

Table 10: Examples of **our augmented annotations** showing the same two original stories from PASTA & SAGA. For outcome descriptions, we use GPT-3.5-Turbo to rewrite the goal descriptions from SAGA as complete sentences describing an ongoing and a completed activity. The condition-outcome relationships are obtained automatically using the SAGA sentence relationships. The outcome label for all outcomes is annotated by an expert based on the story and the outcome relationship is automatically identified based on these annotations.

News Stories	Annotations
Original Story 1: It happens. Just when you think they are gone for good and gather the teddy bears, the little pillow and the tattered blankie and store them in that old trunk in the attic, one of them appears on the doorstep, ready to move back in - with her boyfriend. That' s more or less what happened the other day to Mary Hanford of Salisbury, N. C. Mrs. Hanford, 100, had every reason to figure that her daughter, Elizabeth, was long gone. After all, Elizabeth was 65 and married to a former senator from Kansas named Bob Dole, and they had been living in Washington (in the Watergate, yet) for the better part of 30 years. [National Desk]	**Annotations from the SAGA Extended Version:** Volitional Participant: Elizabeth Elizabeth's Goal: return to her childhood home. **Our annotations:** Outcome described as a completed activity: Elizabeth returned to her childhood home. Outcome Label (Original story): True
Original Story 2: Bond prices for R. H. Macy & Company weakened yesterday on Wall Street speculation that its merchandise shipments might be affected by new credit caution and by a ripple effect from other financially depressed companies. After the markets closed and in response to the widespread speculation about changing credit policies on Macy merchandise, Henry Kassebaum, a senior vice president in New York for the Heller Financial Corporation, a large company that finances merchandise shipments to retailers, said that the company had changed its supplier credit policy in regard to Macy from" revolving credit" to an" order - by - order" policy. In effect, this creates a tighter, more cautious position on granting credit on Macy shipments. Loss Reported Earlier. Macy declined to comment on the possibility of any restriction in its merchandise credit. [Financial Desk]	**Annotations from the SAGA Extended Version:** Participant: Macy Macy's goal: its merchandise shipments might be affected by new credit caution and by a ripple effect from other financially depressed companies. **Our annotations** Outcome described as a completed activity: Macy's merchandise shipments were affected by new credit caution and by a ripple effect from other financially depressed companies. Outcome Label (Original story): True

Table 11: Examples of **our augmented news annotations** with generated conditions and outcome identification. We follow a similar process to that of rewriting SAGA/PASTA annotations, using GPT-3.5-Turbo to rewrite the goal descriptions from SAGA (only describing a completed activity) and obtain expert annotations for outcome labels. We generate conditions using the condition generation process in RQ2 (see §5.2.1).

Type	Prompt
Direct Questions	A:{story} B: {outcome description of ongoing or completed action} For the context in A, Is the statement in B true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.
Imperfective Paradox (No context)	A: {outcome description of past ongoing activity} B: {outcome description of past completed activity} If the statement in A is true, does it necessarily mean that the statement in B is also true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.
Imperfective Paradox (with story)	A: {outcome description of past ongoing activity} B: {outcome description of past completed activity} C: {story} For the context in C, if the statement in A is true, does it necessarily mean that the statement in B is also true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.

Table 12: Prompts used in examining the truth value of outcomes in **RQ1** (§5.1).

Dowty's Imperfective Paradox	Inference
Prompt without context: **A:** Eli was making accurate predictions for the stock market. **B:** Eli made accurate predictions for the stock market. If the statement in A is true, does it necessarily mean that the statement in B is also true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation. **Prompt with story context:** **A:** Eli was making accurate predictions for the stock market. **B:** Eli made accurate predictions for the stock market. **C:** Eli predicted the stock market trends on a lark. When every prediction came true his friends were in awe. They asked him to do it for the next day. His predictions turned out to be sheer luck. Eli's friends were angry when they lost money on their purchases. For the context in C, if the statement in A is true, does it necessarily mean that the statement in B is also true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	Gold Label: Y Generated Label: without context: Y with story context: N
Prompt without context: **A:** Cindy was adopting a puppy. **B:** Cindy adopted a puppy. If the statement in A is true, does it necessarily mean that the statement in B is also true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation. **Prompt with story context:** **A:** Cindy was adopting a puppy. **B:** Cindy adopted a puppy. **C:** Cindy found a cute puppy advertised on facebook. She wanted the puppy so bad. Her husband decided to surprise her. He brought the puppy home to her. Cindy was so happy. For the context in C, if the statement in A is true, does it necessarily mean that the statement in B is also true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	Gold Label: N Generated Label: without context: N with story context: Y

Table 13: Data examples to highlight findings from **RQ1** (§5.1). The two prompts for the imperfective Paradox with and without the story context in Table 12 lead to different inferences when the GPT-4o is used to infer whether Dowty's Imperfective Paradox is true. In these examples, story context influences the model's pragmatic inference.

Direct Questions	Inference
A: Sam got a cold one day. He tried to ignore it. But it grew worse, so he went to the doctor. The doctor told Sam he had the flu! Sam had to take medicine and rest for a week. **B:** Sam was healing from his cold., For the context in A, Is the statement in B true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	Gold Label: Y Generated Label: N
A: Sam got a cold one day. He tried to ignore it. But it grew worse, so he went to the doctor. The doctor told Sam he had the flu! Sam had to take medicine and rest for a week. **B:** Sam healed from his cold. For the context in A, Is the statement in B true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	Gold Label: Y Generated Label: N
A: Jane cooked spinach and chicken for dinner. Her kids hated spinach. They refused to eat it. Jane promised them a new toy if they ate the spinach. Upon hearing this, her kids gobbled up the spinach! **B:** Jane was encouraging her children to eat healthily. For the context in A, Is the statement in B true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	Gold Label: Y Generated Label: N
A: Jane cooked spinach and chicken for dinner. Her kids hated spinach. They refused to eat it. Jane promised them a new toy if they ate the spinach. Upon hearing this, her kids gobbled up the spinach! **B:** Jane encouraged her children to eat healthily. For the context in A, Is the statement in B true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	Gold Label: Y Generated Label: N

Table 14: Data examples to highlight errors with direct questioning discussed in **RQ1** (§5.1). We prompt a model to infer the outcome of both the in-progress and completed action directly. These examples show that the GPT-4o model with the direct prompt from Table 12 generates incorrect labels.

Type	Prompt
Condition Generation	A:{story} B: {outcome} Generate a pair of contrastive conditions relevant to the context in A and the statement in B. Make sure the first condition is supported by the context in A. The conditions should be stative expressions. Do not describe activities.
Identify Condition Single Step	A:{story} B: {outcome} C: {condition1} D: {condition2} For the story in A, is the statement in B true for one of the conditions in C and D and false for the other condition? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.
Identify Condition Standard CoT	A:{story} B: {outcome} C: {condition1} D: {condition2} For the story in A, is the statement in B true for one of the conditions in C and D and false for the other condition? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation. Lets think step by step.
Identify Condition Alternate CoT	A:{story} B: {outcome} C: {condition1} D: {condition2} For the story in A, is the statement in B true for one of the conditions in C and D and false for the other condition? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation. Lets think step by step. We already know the condition in C supports/does not support/is unrelated to the outcome. The condition in D supports/does not support/is unrelated to the outcome. Given this information the answer is
Identify a Condition's Support for the outcome	A:{story} B:{condition} C: {outcome} For the story in A, does the condition in B enable or support the statement in C? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.

Table 15: Prompts used for generating and identifying outcome variant conditions in **RQ2** (§5.2) The first prompt is for generating and the rest are for identifying. Generated outputs from the last prompt for the two conditions are provided as intermediate data for the Alternate CoT.

Prompt	Model Generated Condition Pair & any Issues with the pair
A: Anthony watches a lot of movies about computers. He wished he could impress a venture capitalist to invest in him. After college he designed a website. People were allowed to warn people about other people on this site. A venture capitalist saw the site, and told Anthony he was wasting his time. **B:** Anthony made money from his computer knowledge. Generate a pair of contrastive conditions related to the context in A and the statement in B. Make sure the first condition is supported by the context in A. The conditions should be stative expressions. Do not describe activities.	answer: Anthony was not able to impress a venture capitalist. **[FlanT5-XXL] Model does not generate condition pairs.** Anthony's website was not successful, Anthony's website was successful **[GPT-4o-mini]** Anthony's website was appealing to venture capitalists., Anthony's website was not appealing to venture capitalists. **[GPT-4o] The first condition is not supported by the story as expected.** Anthony was unsuccessful in making money from his computer knowledge, Anthony was successful in making money from his computer knowledge **[Mistral-7BI] conditions are good but lengthy.** The venture capitalist was impressed with Anthony's website, The venture capitalist was unimpressed with Anthony's website. **[Llama-8BI] The first condition is not supported by the story as expected.** Anthony is unsuccessful financially, Anthony is successful financially **[Llama-70BI]**
A: Sam had proposed to Nancy on the steps of the satanic church. Neither was interested in unduly waiting. The wedding date was set for Fall, only a month away. Nancy wore a cream colored dress and laced veil. Sam lifted the veil, kissing her and thinking of the coming honeymoon. **B:** Nancy married Sam as soon as possible. Generate a pair of contrastive conditions related to the context in A and the statement in B. Make sure the first condition is supported by the context in A. The conditions should be stative expressions. Do not describe activities.	answer: The sand was not a good color. The sand was a good color. **[FlanT5-XXL] Unrelated to the story.** Nancy was not ready to marry Sam. Nancy was ready to marry Sam. **[GPT-4o-mini] The first condition is not supported by the story as expected.** The wedding planning was very efficient. Unexpected delays arose, postponing the wedding. **[GPT-4o]** A: The wedding date was not set for immediately., B: The wedding date was set for immediately. **[Mistral-7BI] The first condition is not supported by the story and uses list items.** Nancy is already engaged, Nancy is not married yet. **[Llama-8BI] The conditions are not contrastive.** The wedding was a month away., The wedding was immediate. **[Llama-70BI] The first condition is not supported by the story**

Table 16: Examples of erroneous conditions generated by the various models for the task in §5.2.1.

Type	Prompt
Full Context + No Condition	A:{story} B: {outcome} For the context in A, Is the statement in B true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.
Full Context + Condition	A:{story} B: {condition} C: {outcome} For the context in A, and the condition in B, Is the statement in C true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.
Partial Context + Condition	A:{story without the sentences supporting the condition} B: {condition} C: {outcome} For the context in A, and the condition in B, Is the statement in C true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.
Goal Intent + Condition	A:{Participant} wanted to achieve {goal} B: {condition} C: {outcome} For the context in A, and the condition in B, Is the statement in C true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.

Table 17: Prompts used in examining the truth value of outcomes using conditions with varying context for **RQ3**(§5.3).

Type	Model	Full Story Context		Partial Story Context		Intended Goal		Only Condition
		Baseline #1	+Condition	Baseline #2	+Condition	Baseline #3	+Condition	
Var	FlanT5-XXL	.70 (.88/.78/.44)	**.60** (.89/.80/.12)	.39 (.68/.35/.13)	.53 (.79/.62/.18)	.30 (.74/.11/.05)	.50 (.80/.65/.06)	.47 (.59/.70/.12)
	GPT-4o-mini	.66 (.87/.82/.29)	**.61** (.89/.84/.10)	.50 (.70/.64/.16)	.54 (.81/.70/.09)	.29 (.49/.26/.10)	.47 (.77/.61/.02)	.40 (.54/.55/.10)
	GPT-4o	.67 (.84/.83/.33)	.66 (.86/.82/.30)	.35 (.42/.50/.13)	.48 (.62/.66/.15)	.05 (.02/.03/.09)	.29 (.40/.39/.12)	.13 (.07/.22/.09)
	Mistral-7BI	.53 (.83/.76/.00)	**.59** (.83/.78/.18)	.42 (.60/.46/.19)	.49 (.64/.65/.18)	.30 (.67/.14/.09)	.40 (.53/.58/.10)	.37 (.41/.60/.09)
	Llama-8BI	.53 (.83/.76/.00)	**.54** (.84/.76/.00)	.36 (.64/.42/.00)	.47 (.74/.68/.00)	.32 (.58/.39/.00)	.47 (.74/.68/.00)	.44 (.63/.63/.06)
	Llama-70BI	.62 (.87/.79/.22)	**.63** (.87/.80/.22)	.39 (.55/.48/.12)	.53 (.67/.72/.19)	.23 (21/.46/.02)	.46 (.60/.69/.08)	.37 (.47/.57/.07)
InVar	FlanT5-XXL	.52 (.82/.62/.12)	.50 (.82/.66/.03)	.47 (.75/.44/.21)	.47 (.77/.54/.12)	.28 (.74/.06/.03)	.43 (.72/.50/.08)	.39 (.43/.50/.25)
	GPT-4o-mini	.58 (.81/.69/.23)	.53 (.80/.66/.14)	.43 (.62/.45/.23)	.46 (.74/.45/.19)	.33 (.54/.31/.15)	.41 (.66/.41/.15)	.33 (.42/.32/.26)
	GPT-4o	.68 (.80/.72/.52)	.61 (.78/.62/.42)	.31 (.29/.36/.28)	.39 (.49/.40/.28)	.12 (.08/.03/.26)	.27 (.33/.22/.26)	.15 (.07/.11/.26)
	Mistral-7BI	.49 (.77/.63/.06)	.51 (.76/.61/.15)	.43 (.72/.51/.05)	.45 (.64/.47/.24)	.37 (.69/.23/.20)	.36 (.43/.38/.27)	.30 (.29/.39/.24)
	Llama-8BI	.49 (.81/.66/.00)	.44 (.73/.59/.00)	.41 (.72/.51/.00)	.41 (.68/.54/.00)	.34 (.61/.40/.00)	.37 (.63/.47/.00)	.34 (.53/.46/.04)
	Llama-70BI	.54 (.78/.63/.20)	.54 (.77/.64/.21)	.46 (.71/.51/.16)	.48 (.69/.59/.16)	.22 (.18/.44/.04)	.38 (.46/.46/.21)	.33 (.33/.37/.29)
All	FlanT5-XXL	.58 (.85/.69/.19)	.54 (.85/.72/.05)	.43 (.72/.40/.18)	.49 (.77/.57/.14)	.29 (.74/.09/.03)	.46 (.75/.56/.08)	.43 (.50/.50/.21)
	GPT-4o-mini	.61 (.83/.75/.24)	.57 (.84/.74/.13)	.47 (.65/.54/.21)	.50 (.77/.57/.16)	.31 (.52/.29/.13)	.44 (.70/.50/.11)	.37 (.42/.43/.21)
	GPT-4o	**.69** (.81/.77/.47)	.64 (.81/.71/.39)	.33 (.34/.43/.23)	.44 (.55/.53/.24)	.09 (.05/.03/.19)	.29 (.36/.30/.21)	.14 (.07/.16/.20)
	Mistral-7BI	.51 (.80/.69/.05)	.54 (.78/.68/.16)	.42 (.68/.49/.10)	.47 (.64/.55/.22)	.34 (.68/.19/.16)	.39 (.47/.47/.22)	.33 (.34/.48/.19)
	Llama-8BI	.51 (.82/.70/.00)	.48 (.78/.66/.00)	.39 (.69/.47/.00)	.43 (.70/.60/.00)	.33 (.60/.40/.00)	.41 (.67/.56/.00)	.38 (.57/.53/.05)
	Llama-70BI	.57 (.81/.70/.21)	.58 (.81/.71/.22)	.43 (.65/.50/.14)	.50 (.68/.65/.17)	.22 (.19/.45/.03)	.42 .52/.57/.17)	.36 (.39/.46/.22)

Table 18: Outcome validation performance (Macro F1 and individual F1 scores of True/False/Unsure labels when **using outcome-variant & invariant conditions**. The 'All' type shows F1 when both condition types are used together. We compare the 4 varying contexts listed in §5.3 and 3 baselines consisting of only the story contexts – full, partial and goal. **Bolded** numbers show variant conditions lead to 2-8 % improvement in model performance (except for GPT-4o & GPT-4o-mini) over the 'All' baseline #1.

Model	Full Story Baseline	Varying Context With Condition		Only Condition
		Full Story	Intended Goal	
FlanT5-XXL	.57 (.82/.71/.19)	.55 (.80/.69/.14)	.38 (.66/.38/.09)	.33 (.45/.36/.18)
GPT-4o-mini	.60 (.81/.77/.21)	.51 (.74/.70/.10)	.36 (.59/.37/.11)	.34 (.41/.38/.23)
GPT-4o	.69 (.81/.79/.46)	.59 (.72/.70/.35)	.30 (.39/.32/.20)	.16 (.10/.19/.19)
Mistral-7BI	.52 (.76/.71/.09)	.49 (.69/.66/.12)	.38 (.39/.53/.21)	.33 (.36/.53/.09)
Llama-8BI	.49 (.77/.70/.00)	.45 (.68/.66/.00)	.36 (.55/.52/.00)	.34 (.51/.52/.00)
Llama-70BI	.55 (.78/.73/.13)	.52 (.73/.69/.15)	.37(.42/.57/.13)	.38 (.41/.54/.19)

Table 19: Outcome validation performance (F1 scores of True/False/Unsure labels) **using generated conditions**. We compare the 4 varying contexts leaving out the Partial Story context as we do not identify which sentences the condition depends upon. see §5.3 for the task setup. We could not generate conditions for some stories and performance with generated conditions is lower than with annotated conditions for all models.

Prompt	Model generations & Issues
A: The proposal was unsuccessful. **B:** Danny proposed to Beth. For the condition in A, is the statement in B true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	**FlanT5-XXL** incorrectly generates an N when only the condition is provided but correctly generates a Y when story context is availablen as in the next row. **GPT-4o-mini** correctly generates Y for all settings where the condition is provided and incorrectly generates a N otherwise **GPT-4o** generates a U for all contexts where the condition is provided and correctly generates a Y with story context. **Mistral7BI** incorrectly generates a N when the condition is provided but correctly generates a Y with story context. **Llama8BI** correctly generates a Y for all settings where the condition is provided. **Llama70BI** incorrectly generates N because the condition states something that happened later. For the settings where some story context is available the model correctly generates Y.
A: Danny got down on one knee. He asked Beth to marry him. Beth felt very awkward. She didn't want to marry Danny! To be honest, she told him she did not want to marry him. **B:** Beth was brutally honest. **C:** Beth let Danny down gently. For the context in A and the condition in B, Is the statement in C true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	**FlanT5-XXL** correctly generates a Y when provided with story context and generates a N or U otherwise. **GPT-4o-mini** correctly generates a Y for all settings where the story context is provided and generates a U otherwise. **GPT-4o** correctly generates a Y for all settings where the story context is provided and generates a U otherwise. **Mistral7BI** correctly generates a Y for all settings where the story context is provided and incorrectly generates a N otherwise. **Llama8BI** correctly generates Y and generates N for all settings where the condition is provided. **Llama70BI** correctly generates Y whereas it incorrectly generates N when only the context is provided.

Table 20: Examples of **outcome identification errors** from the various settings in §5.3 using Model generated conditions. We show 2 prompts with one not using a story context and the other using a story context. We show how models behave for various settings using a single story and 2 outcome descriptions for the two participants.

Prompt	Model Generations & Issues
A: When George Bush declared last June that he would not raise taxes, skeptical reporters gave him repeated chances to hedge. Could he support higher excise taxes ? No, Mr. Bush said. Would he use" revenue enhancers" ? No, Mr. Bush insisted. **B:** George Herbert Walker Bush did not raise taxes Generate a pair of contrastive conditions related to the context in A and the statement in B. Make sure the first condition is supported by the context in A. The conditions should be stative expressions. Do not describe activities.	answer: George Bush raised taxes. **[FlanT5-XXL]** George Bush is evasive., George Bush is straightforward. **[GPT-4o-mini]** George Bush is committed to his promise., George Bush is open to changing his stance.**[GPT-4o]** George Herbert Walker Bush did not commit to not raising taxes., George Herbert Walker Bush committed to not raising taxes.**[Mistral-7BI]** George Bush had a plan to increase taxes., George Bush did not have a plan to increase taxes.**[Llama-8BI]** George Bush was open to new taxes. George Bush was not open to new taxes. **[Llama-70BI]**
A: When George Bush declared last June that he would not raise taxes, skeptical reporters gave him repeated chances to hedge. Could he support higher excise taxes ? No, Mr. Bush said. Would he use "revenue enhancers" ? No, Mr. Bush insisted. **B:** {generated condition (the first one) from the above row} **C:** George Herbert Walker Bush did not raise taxes For the context in A and the condition in B, Is the statement in C true? Please indicate with a 'Y' for yes, 'N' for no and 'U' for unsure. Do not give an explanation.	**FlanT5-XXL** generates correctly a Y for all contexts containing the story or the George Bush's intended goal and generate N for the condition only setting. **GPT-4o-mini** correctly generates a Y when the condition is not part of the context but generates a N when it is. **GPT-4o** generates a U for most contexts and a N when the condition provides the most context. **Mistral-7BI** correctly generates a Y when the condition is not part of the context but generates a N when it is. **Llama8BI** incorrectly generates a N for all contexts using its prior knowledge **Llama70BI** incorrectly generates a N for all contexts using its prior knowledge

Table 21: Examples of **News story** condition and outcome generations which are **biased due to models' knowledge of past news events**. All Models except FlanT5-XXL and Mistral make use of their knowledge of prior events when generating conditions and identifying outcome labels. FlanT5-XXL does not generate condition pairs but uses its available context; Mistral-7BI generates correct but long condition pairs and considers all available context.

Model	Full Story Baseline	Varying Context With Condition		
		Full Story	Intended Goal	Only Condition
FlanT5-XXL	.54 (.78/.44/.38)	.58 (.74/.41/.40)	.34 (.63/.20/.19)	.35 (.42/.31/.32)
GPT-4o-mini	.51 (.75/.50/.28)	.46 (.72/.42/.23)	.29 (.54/.30/.05)	.32 (.45/.32/.21)
GPT-4o	.60 (.68/.55/.56)	.58 (.69/.52/.53)	.34 (34/.33/.35)	.31 (.27/.32/.33)
Mistral-7BI	.40 (.74/.44/.03)	.35 (.54/.40/.12)	.34 (.33/.38/.30)	.24 (.19/.35/.18)
Llama-8BI	.39 (.73/.44/.00)	.39 (.71/.45/.00)	.31 (56/.37/.00)	.26 (.44/.34/.00)
Llama-70BI	.37 (.64/.40/.05)	.34 (.60/.41/.03)	.28 (.46/.39/.00)	.29 (.43/.38/.06)

Table 22: **News Story** Outcome validation performance (Macro F1 and F1 scores of True/False/Unsure labels in parenthesis) **using generated conditions**. We compare the 4 varying contexts leaving out the Partial Story context as we do not identify which sentences the condition depends upon. See §6 for the task details.

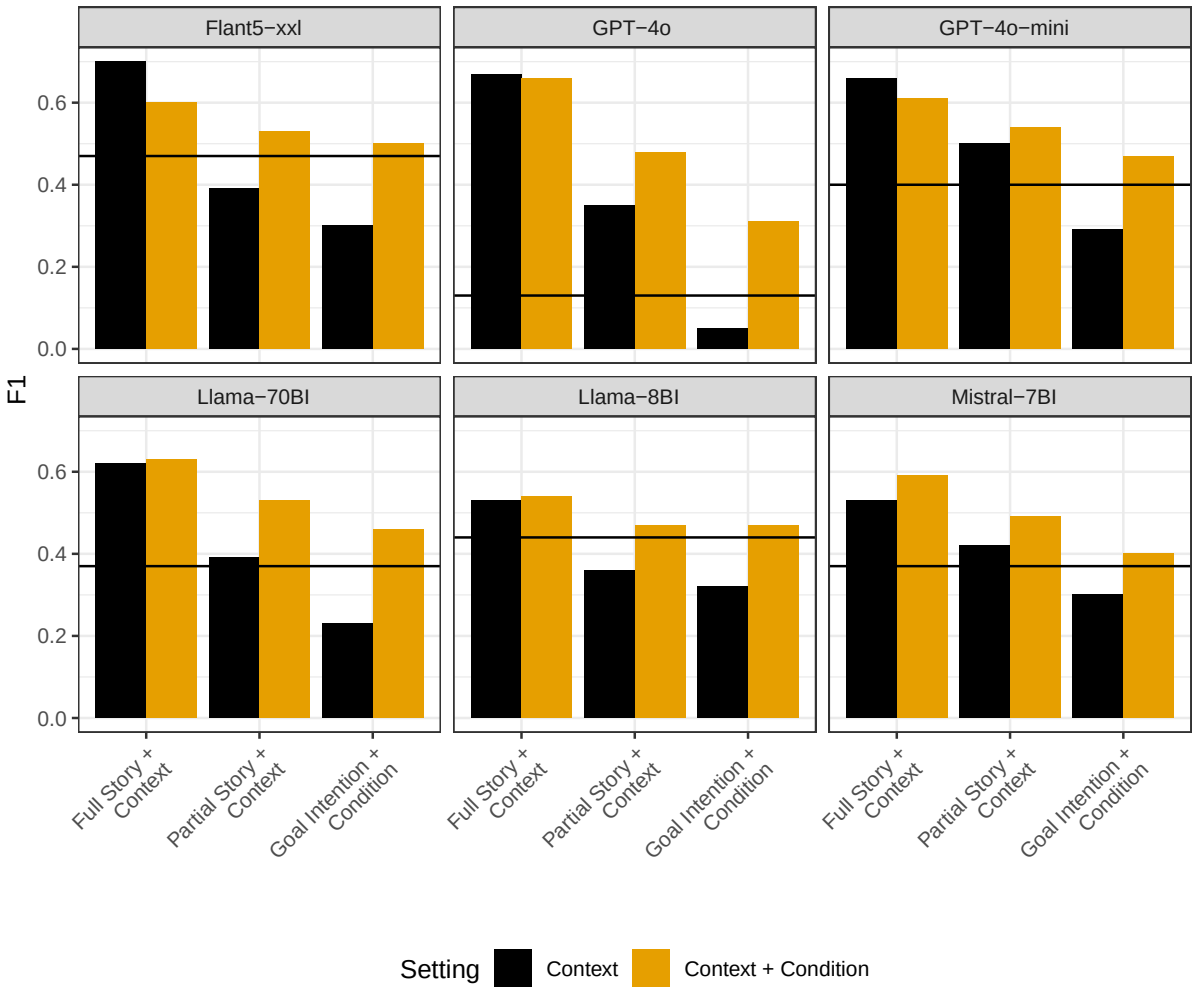


Figure 7: Models are able to utilize conditions in addition to the story context to improve performance when identifying outcome labels. The exceptions are FlanT5-XXL, GPT-4o-mini and GPT-4o which drop in performance when using conditions with full story context. The black bar shows performance (Macro F1) with story context alone and the yellow bar shows the improvement when annotated variant conditions are used in addition to the story contexts. The black line indicates the performance of the condition only task setting.