Tricking Conditional LLMs

Mike Ledford
Department of Computer Science
University of Maryland, College Park
mledfor@umd.edu

Nate Kadawedduwa
Department of Computer Science
University of Maryland, College Park
nkadawed@terpmail.umd.edu

Abstract

Pre-trained Conditional Large Language Models (LLMs) have shown exceptional performance in various NLP tasks, yet they remain susceptible to generating erroneous content due to their reliance on patterns in data rather than true world knowledge. This study proposes a systematic framework for intentionally misleading LLMs through prompt engineering. By evaluating models of different sizes, we find that larger LLMs struggle with prompts resembling common training data patterns, requiring multi-hop reasoning, or combining complex domains like commonsense reasoning and temporal discontinuity. Using a hybrid few-shot and zero-shot learning approach, our analysis reveals that these high-level challenges hinder the models' ability to reproduce similar prompts effectively, suggesting that such themes represent conceptual boundaries that current LLMs struggle to integrate or comprehend.

1 Introduction

Generative language models (LMs) all rely on some representation of real-world knowledge to create narratives or accomplish reasoning tasks. Whether this is in the form of an explicit knowledge graph or embedded in the model's parameters, its limitations are always tied to the breadth of its training data.

The symptoms of limited implicit knowledge can show up in cases where the distribution of probabilities in the local learned space do not match with their occurrences in real life. For example, a prompt like "Jo put a lasagna in the oven but forgot to turn it on. When she opened the oven door..." may lead the model to continue a story about a burnt lasagna rather than a cold, un-cooked one because opening oven doors is more frequently linked to burnt food in the training data regardless of the physical reactions.

This issue often arises because large language models (LLMs) like GPT are trained to predict the next word based on data patterns rather than a genuine understanding of the real world. A primary contributor to this problem is reporting bias, which systematically distorts the frequency of certain events, outcomes, and properties in text (Gordon and Van Durme, 2013). Such misunderstandings were far more common in GPT-3 than they are in GPT-4 (Achiam et al., 2023). With vastly more parameters and a significantly larger training dataset, modern LLMs are now much more resilient to being misled. However, these state-of-the-art models still suffer from inherent biases in training data leading to ambiguous, erroneous, and biased outputs (Raiaan et al., 2024).

While success stories present a picture of LLMs' impressive capabilities, it is equally crucial to highlight its pitfalls to understand the full spectrum of their impact. In this work, we seek to investigate systematic ways to lead conditional LLMs astray. Identifying such ways could help determine whether errors arise from issues such as coreference resolution, commonsense reasoning, or contextual ambiguity at lexical, syntactic, or semantic levels, along with other underlying phenomena. Consequently, these insights could inform future design and training strategies to enhance the robustness of language models as they become increasingly versatile and widely used across NLP applications.

2 Systematic Approach

To systematically explore ways to lead conditional LLMs astray, we executed a four-phase approach.

Problem Identification — This phase involved systematically exploring and evaluating a range of challenging prompts from datasets and benchmarks known to challenge current state-of-the-art models. The goal was to curate a new set of human-generated challenging prompts designed to potentially mislead the LLM.

Prompt	Observations
Alex is on a game show and he's given the choice of three doors: Behind one door is a new car; behind the others, donkeys. He picks a door, say No.1, and the host asks you “Do you want to pick door No.2 instead?” In order to maximize his chances of winning the car, Alex...	The prompt leads the model to the common Monty Hall problem which actually does not apply here.

Figure 1: An example prompt resembling a common logic puzzle (Monty Hall Problem) except with slight modifications to the narrative

Trend Analysis — This phased involved assessing state-of-the-art models on the curated set of human-generated prompts. Quantitative and qualitative analyses are performed to evaluate model performance for each prompt, with the primary objective of identifying and categorizing syntactic patterns that consistently affect model behavior.

Concept Verification — This phase involved ensuring that the syntactic patterns identified in the *Trend Analysis* phase were distinct from any artifacts specific to individual models. The objectives were twofold: (1) to create additional human-generated examples featuring these syntactic patterns and evaluate model performance on them, and (2) to leverage LLMs to generate examples with these syntactic patterns, examining whether models exhibit higher accuracy on examples generated by the LLMs based on similar provided instances.

Experimentation with Teaching — This phase involved exploring the viability of In-Context Learning (ICL) to teach models to address challenges posed by prompts containing the syntactic patterns identified during the *Concept Verification* phase.

Although the remainder of this paper does not delve deeply into each of these phases, it is structured to discuss dataset creation, the experiments conducted using the dataset, analysis of the experimental results, and finally, concept verification and initial experimentation with ICL.

3 Dataset

Here, we present our dataset, which comprises 10 new carefully crafted, human-generated prompts designed to challenge models by requiring the integration of world knowledge, contextual reasoning, and common sense. These prompts are specifically aimed at evaluating a model's reasoning capabilities, with a focus on determining whether it can process natural language and reason about the prompts in a human-like manner. By meticulously designing these prompts, we aim to systematically expose potential factors that contribute to LLMs being misled. Such factors could include vulnerabilities like reliance on model-specific artifacts, over-dependence on statistical patterns inherent to the training data, and challenges in coreference resolution, among others.

Prompt Construction. We adopted a systematic approach to design prompts that required the model to generate either the conclusion of a narrative or the answer to a logic question. We began with an example prompt, *"Jo put a lasagna in the oven but forgot to turn it on. When she opened the oven door..."*, and crafted additional prompts with similar semantic structures. These prompts featured agents (characters) performing intentional or unintentional actions connected by causal or temporal relationships. Each prompt was designed to emphasize semantic interactions across multiple lines, creating a coherent narrative progression that culminated in an open-ended scenario ("..."), where the model was tasked with continuing the story.

To expand our list of prompts, we conducted a comprehensive literature review and analyzed datasets specifically designed to challenge state-of-the-art models, including WINOGRAD, PIQA, COMMONSENSEQA, COM2SENSE, and LOGIQA (Sakaguchi et al., 2021; Bisk et al., 2020; Talmor et al., 2018; Singh et al., 2021; Liu et al., 2020). These datasets served as inspiration, as each presents unique challenges that require models to resolve ambiguities through world knowledge, contextual reasoning, and common sense. Additionally, we drew from the work of Williams and Huckle (2024), which highlights "easy" problems that LLMs fail to solve, providing valuable insights into systematic model errors.

Prompt Description. For each prompt, we provided detailed annotations capturing key observations to evaluate whether the model was misled. Figure 1 illustrates one such prompt. This particular example is a modified version of the Monty Hall problem, where the probabilities remain unchanged. If the model responds with the Monty Hall solution (e.g., switching to door No. 2 to increase its probability of winning to $2/3$), we can conclude that the model was misled. Subsequently, we proceed to identify the underlying root cause of this error. A comprehensive list of prompts and

Prompt #	Prompt	Intended Behavior	Observations (Commonsense reasoning domains, coreference resolution, ambiguous language, etc)	Models Tricked	Models NOT tricked	% Percentage Tricked
1	John woke up late to work and was in a rush to get into the office. Yesterday, he worked on a house project and forgot to move a large piece of plywood with nails sticking out of it from the ground behind his car. In a rush to leave, he quickly pulled forward out of his parking spot and drove to work. After 8 hours of working, he returned to the parking lot and ...	The model should allude to the fact that the car is fine and there isnt a flat tire because he drove forward and not in reverse thus avoiding the nails behind him	1. Social Commonsense (late for work, he's most likely going to speed to work) 2. Causal Commonsense (if he reverses, he runs over plank and has a nail in his tire) 3. Physical Commonsense (nail + tire = flat tire) 4. Temporal Commonsense (if he has a nail in his tire, after 8 hours of work, the tire is most likely losing air)	ALL	ALL	100%
2	Charley took Amy to the state fair but forgot that she is afraid of heights. This year's state fair has two rides, a carousel and a drop tower. Before eating lunch, Charley and Amy decided to go on a ride. Charley and Amy rode the ...	The model should allude to the fact that Charley and Amy went to the carousel because it is much lower to the grouund than drop tower	1. Physical Commonsense (drop tower is higher than a carousel) 2. Social Commonsense (Someone afraid of heights will most likely not go on a high ride)	NONE	NONE	0%
3	Anamarie went to go watch her play little league baseball at the local park. His team was down by 3 runs in the bottom of the last inning. Her son was the second to last batter. The current batter hit a grand slam..	The model should allude to the fact that the game is over because a grand slam would put the team ahead and no need for Anamarie's son to bat	1. Social Commonsense (grand slam = 4 runs) 2. Causal commonsense (down by 3 and scoring 4 runs means the game is over)	LLama 3.2-3B Instruct Claude 3.5 Sonnet Sonar Huge	GPT-4	75%
4	Tina and her family went to Copper mountain to go skiing. Tina went to the freezer outside to retrieve the chicken that was placed in there yesterday. On her way back inside, she stopped outside to take pictures of the snow. She placed the chicken on the chair covered in snow to take pictures. After 20 minutes, Tina and her family changed and went skiing for 5 hours. When she came back to cook the chicken...	Interesting to see if the model thinks the chicken thawed or not which affects whether it is cookable or not	1. Physical commonsense 2. Causal commonsense (leaving chicken out (cause) leads to potentially spoilage (effects) 3. Temporal commonsense (spoilage could occur after a set # of hours)	LLama 3.2-3B Instruct	GPT-4 Claude 3.5 Sonnet* Sonar Huge*	25%
5	Alex is on a game show and he's given the choice of three doors: Behind one door is a new car; behind the others, donkeys. He picks a door, say No.1, and the host asks you "Do you want to pick door No.2 instead?" In order to maximize his chances of winning the car, Alex...	The prompt leads the model to the common Monty Hall problem which actually does not apply here.	1. Causal commonsense (probability does not change)	GPT-4 Claude 3.5 Sonnet Sonar Huge	LLama 3.2-3B Instruct	75%
6	In a toy box, there's a red ball, a blue truck, and a green dinosaur. The red ball is not next to the blue truck, and the green dinosaur is next to the red ball. The toy in the middle is...	The model needs to understand the logical physical relationships between objects.	1. Physical (Spatial) Commonsense	GPT-4 Claude 3.5 Sonnet	LLama 3.2-3B Instruct Sonar Huge	50%
7	Josh was sipping from his fancy, shiny wine glass made of plastic when, suddenly, his brother ran into the room and deliverd the horrifying news. their parents had been in a terrible car accident. Josh dropped the glass and...	The plot between the family could distract from the simple logical reasoning of the glass being made of rubber and therefore not shattering.	1. Physical Commonsense	GPT-4 Sonar Huge	LLama 3.2-3B Instruct Claude 3.5 Sonnet	50%
8	George wanted to measure exactly 4 gallons of water with only a 3-gallon, 5-gallon, and 4-gallon jug so he...	This prompt is similar to the jug puzzle but has an easier answer if the model recognizes the simple connection between the amount of water and size of the jugs	1. Physical Commonsense (simply recognize matching units of measure)	LLama 3.2-3B Instruct Sonar Huge	GPT-4 Claude 3.5 Sonnet	50%
9	Johnny was playing a game of darts with his college roommates. He needed a bullseye to win the game. When Johnny through the rubber dart...	The dart should not stick to the board because it is rubber.	1. Physical Commonsense (properties of a rubber dart)	ALL	ALL	100%
10	Tina was on her way to the movies when stopped by to pickup her friend stacy. Stacy opened the passenger door and got in the front seat. Because Tina takes her kids to school, she keeps the child lock on. When Tina and Stacy got to the parking lot and parked...	Stacy should be able to get out because child lock only applies to the seats in the back	1. Social Commonsense 2. Coreference Resolution	GPT-4 Claude 3.5 Sonnet Sonar Huge	LLama 3.2-3B Instruct	75%

Figure 2: Evaluation of each prompt across all four models. Models performance for each prompt varied. One prompt tricked all four models and one prompt did not trick any models.

their corresponding observations is included in Appendix A.1.

4 Experiments

We summarize our experimental set up, models, and the evaluation strategy used in this work.

4.1 Experimental Setup

In this work, we test four pre-trained conditional LLMs on our curated set of prompts and report their performance. The LLMs included three large-scale models—GPT-4o from OpenAI, Sonar Huge from Perplexity, and Claude 3.5 Sonnet from Anthropic—as well as a smaller, open-source model, Llama-3.2-3B-Instruct, developed by Meta. We specifically selected the instruction-tuned variant of Llama-3.2-3B over the base model due to concerns about the reliability and robustness of the outputs from the latter. Importantly, we intentionally included models of varying sizes to enable a comparative analysis of smaller models versus large-scale systems, thereby enhancing the generalizability of our findings. All models were tested in a zero-shot setting with full precision (non-quantized).

4.2 Evaluation Strategy

We evaluate the performance of the models on the curated prompts using a structured methodology and conduct a comprehensive comparative analysis for each prompt. For every prompt, we quantitatively calculate the number of models that were *tricked* and *not tricked*, using a binary assessment (yes/no). This evaluation is grounded in human judgment, focusing on the coherence of the responses and determining whether the predicted events generated by the models are plausible or logically consistent, particularly for logic-based questions. Additionally, we examine whether the models relied on statistical patterns rather than reasoning to generate their responses. The proportion of models that were tricked by a prompt serves as an accuracy metric for our evaluation.

Following the quantitative analysis, we qualitatively assess each prompt, annotating semantic and syntactic patterns as well as other factors that might have misled the models. For example, we consider whether the prompt involved multiple domains of commonsense reasoning, parallel structures with opposing actions, ambiguity introduced by coreference resolution, or a combination of multiple concepts or prepositions. This dual quantitative and qualitative approach ensures a robust understanding of the factors influencing model performance.

5 Results and Discussion

Figure 2 presents the percentage of models tricked by each prompt alongside a qualitative assessment of the syntactic patterns and factors that may have contributed to the models being misled. Notably, Prompts 1 and 9 (100%) and Prompts 3, 9 and 5 (75%) had the highest percentages of models tricked, while Prompts 2 (0%) and 4 (25%) recorded the lowest percentages. These findings highlight the varying levels of difficulty among the prompts and their ability to exploit specific model vulnerabilities. Based on our findings, the most ef-

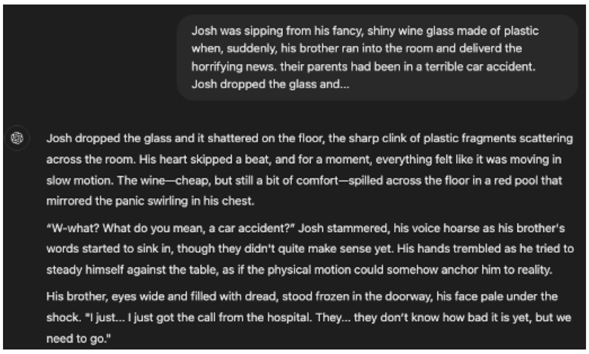


Figure 3: Common occurrence of events in training data leads to incorrect model response

fective prompts for misleading models consistently exhibited the following three prompt sensitivity themes:

Theme 1: Strongly resemble common occurrences in the training data. Prompts with phrasing resembling common logic puzzles or incorporating familiar action-event pairs (e.g., "dropping a wine glass causes the glass to shatter") often elicited predictable responses from the models. In such cases, the models typically provided the solution to the original logic puzzle, failing to recognize that the puzzle had been modified. Similarly, the models often predicted the most statistically probable outcome of an action, disregarding modifications to the underlying physics or characteristics of the action specified in the prompt. This is evident in Figure 3, where GPT-4o incorrectly assumes that a wine glass *made of plastic* shatters when dropped. In contrast, human intuition recognizes that most plastics are durable and resistant to breaking, highlighting a gap in the model's reasoning capabilities.

Theme 2: Multi-hop reasoning with multiple concepts and prepositions. Models consistently struggled when required to draw conclusions by synthesizing multiple pieces of information from different parts of the input. Performance typically declined as the number of concepts and prepositions in the prompts increased. See Table 1 for an example. Furthermore, the models exhibited particularly poor performance on prompts involving qualitative spatial reasoning, where they were required to deduce spatial orientation and motion relative to a reference point without explicit mathematical representations (e.g., "in front" instead of "2 meters north"). When these features were combined, LLMs demonstrated significant difficulty in maintaining context across long or complex prompts. See Appendix A.2 for another example.

# of Objects	# of Prepositions	Tricked? Rate
3	2	$4/5 = 80\%$
3	3	$2/5 = 40\%$
3	4	$2/5 = 40\%$
3	5	$1/5 = 20\%$

Table 1: Prompt: *In a toy box, there's a* ***[List of objects]. [List of prepositions]****. The toy in the middle is...*[1]

Theme 3: Increased complexity and cognitive load with multi-line prompts. It is well-established that the cognitive load on LLMs increases with multi-line prompts, making accurate processing and response generation more challenging. Our observations revealed mixed performance across prompts that varied in the number and types of commonsense reasoning domains. Notably, the majority of prompts that successfully tricked the models required reasoning across at least two domains of commonsense reasoning, with one domain consistently being causal reasoning. This finding underscores the persistent challenges that LLMs face in causal commonsense reasoning tasks. Additionally, prompts featuring inherent temporal discontinuity further impacted performance, as models struggled to determine the relevance of events occurring on different days (e.g., the day before and the current day). These results highlight specific areas where LLMs still exhibit significant limitations.

5.1 Concept Verification

In this section, we present the results of testing models on both human-generated and LLM-generated prompts that reflect the three identified themes of prompt sensitivity. The primary goal was to confirm that the observed themes are independent of any model-specific artifacts or situational cues. Additionally, we explored the feasibility of using an LLM as a pipeline to efficiently generate supplementary examples aligned with these themes. Through this pipeline, we examined whether a model demonstrates higher accuracy on examples it generates itself when provided with an initial reference example.

The results of testing a model on a new human-generated prompt reflecting these themes are presented in Figure 4. In this example, the model fails

[1]Limited trials are for demonstration and are not conclusive. Tested on Llama 3.2-3B.

to reason visually about the spatial orientation of the *shelf* relative to the car (e.g., in front) and the directional motion of *the car reversing*. Instead, it incorrectly assumes that the car reverses into the shelf. This demonstrates the model's limitations in handling directional and spatial reasoning, leading to misalignment in its predictions.

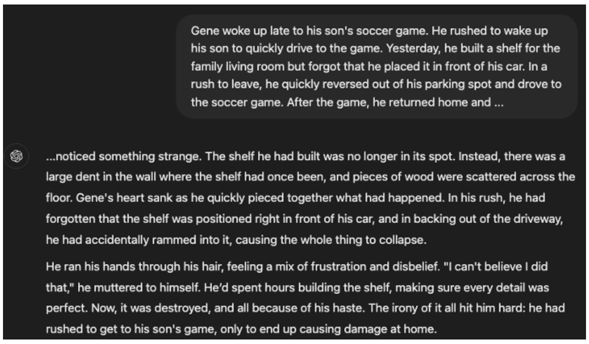


Figure 4: GPT-4o generates incorrect response due to qualitative spatial reasoning, multiple commonsense domains, and temporal discontinuity in the prompt.

5.2 Model Generated Prompts

In further testing, we explore how proficient Llama (Instruct) or our other models are at generating prompts "similar" to our dataset. We consider similar prompts to be those which implement one or more of our above themes in the same partial story/logic puzzle formats. In regards to the story format, it typically includes a sentence which establishes background and finishes with a corresponding implication which denies the most common occurrence (i.e., introduce a plastic wine glass; implication of dropping it...).

One approach to generating similar prompts was through Few Shot Learning (FSL). By providing our model with the best (highest trick rate) prompts from our dataset and constructing a prompt asking for output of the same structure, we experiment to see if it can pick up on the high-level themes which we outlined (Appendix A.3).

Another approach is through Zero Shot Learning (ZSL) with explicitly defined themes. Rather than giving examples and assuming that the model will pick up on the similarities (Effective for low-level patterns but more difficult for complex ones), we explore if outlining the themes in the prompt will produce better results. A hybrid method which combines the two above was also tested (Appendix A.4).

The results for all of these methods did either fail to trick all tested LLMs 100% of the time or did not have a clear condition that would indicate if the model was tricked or not. Further prompt engineering was conducted by producing a full prompt and manually inserting ellipses after the first clause in the first non-exposition sentence. This did not yield any further improvements (Appendix A.5)

From these results, we can extrapolate that our defined themes were indeed difficult for LLMs to handle because they could not easily reproduce them. The frequency with which these LLMs produced a standard story with no twist or a subversion with no previous set-up, suggests that these high-level concepts are too difficult for the models to integrate or parse.

5.3 Future Work

Through experimentation and evaluation, we have outlined a systematic method and set of themes through which Llama 3.2 and some modern LLMs can be consistently tricked. The next logical step which we sought to at least layout is training models to better deal with these difficult themes. With limited testing, we propose a fine-tuning method as the best approach for this task. While in-context learning was considered, it has shown itself to be inconsistent when dealing with such complex topics or anything beyond surface level structure.

The model generated prompts from our study may not have been effective enough to set up a pipeline for automatically generating similar prompts, but a hybrid method combining FSL and ZSL, along with some more involved prompt engineering and larger datasets, could possibly see some success.

Acknowledgments

We would like to thank Professor Rudinger and Shramay for their valuable feedback in the creation and proposal of this project.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. Com2sense: A commonsense reasoning benchmark with complementary sentences. *arXiv preprint arXiv:2106.00969*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Sean Williams and James Huckle. 2024. Easy problems that llms get wrong. *arXiv preprint arXiv:2405.19616*.

A Appendix

A.1 Full List of our 10 Human Generated Prompts

1. John woke up late to work and was in a rush to get into the office. Yesterday, he worked on a house project and forgot to move a large piece of plywood with nails sticking out of it from the ground behind his car. In a rush to leave, he quickly pulled forward out of his parking spot and drove to work. After 8 hours of working, he returned to the parking lot ...
   Observation: The model should allude to the fact that the car is fine and there isn't a flat tire because he drove forward and not in reverse thus avoiding the nails behind him

2. Charley took Amy to the state fair but forgot that she is afraid of heights. This year's state fair has two rides, a carousel and a drop tower. Before eating lunch, Charley and Amy decided to go on a ride. Charley and Amy rode the ...
   Observation: The model should allude to the fact that Charley and Amy went to the carousel because it is much lower to the ground then drop tower

3. Anamarie went to go watch her play little league baseball at the local park. His team was down by 3 runs in the bottom of the last inning. Her son was the second to last batter. The current batter hit a grand slamv...
   Observation: The model should allude to the fact that the game is over because a grand slam would put the team ahead and no need for Anamarie's son to bat

4. Tina and her family went to Copper mountain to go skiing. Tina went to the freezer outside to retrieve the chicken that was placed in there yesterday. On her way back inside, she stopped outside to take pictures of the snow. She placed the chicken on the chair covered in snow to take pictures. After 20 minutes, Tina and her family changed and went skiing for 5 hours. When she came back to cook the chicken ...
   Observation: Interesting to see if the model thinks the chicken thawed or not which affects whether it is cookable or not

5. Alex is on a game show and he's given the choice of three doors: Behind one door is a new car; behind the others, donkeys. He picks a door, say No.1, and the host asks you "Do you want to pick door No.2 instead?" In order to maximize his chances of winning the car, Alex ...
   Observation: The prompt leads the model to the common Monty Hall problem which actually does not apply here.

6. In a toy box, there's a red ball, a blue truck, and a green dinosaur. The red ball is not next to the blue truck, and the green dinosaur is next to the red ball. The toy in the middle is ...
   Observation: The model needs to understand the logical physical relationships between objects.

7. Josh was sipping from his fancy, shiny wine glass made of plastic when, suddenly, his

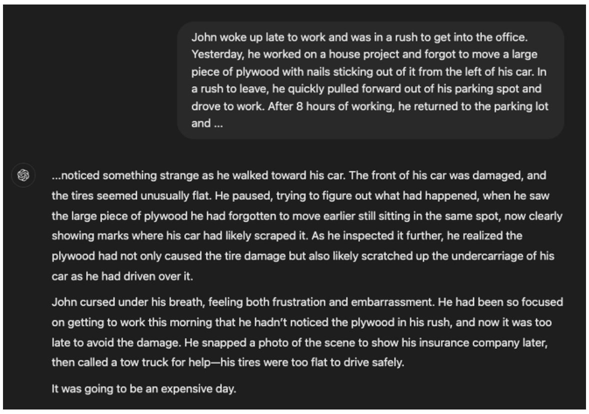
brother ran into the room and delivered the horrifying news. their parents had been in a terrible car accident. Josh dropped the glass and ...
Observation: The plot between the family could distract from the simple logical reasoning of the glass being made of rubber and therefore not shattering.

8. George wanted to measure exactly 4 gallons of water with only a 3-gallon, 5-gallon, and 4-gallon jug so he ...
Observation: This prompt is similar to the jug puzzle but has an easier answer if the model recognizes the simple connection between the amount of water and size of the jugs

9. Johnny was playing a game of darts with his college roommates. He needed a bullseye to win the game. When Johnny through the rubber dart ...
Observation: The dart should not stick to the board because it is rubber.

10. Tina was on her way to the movies when stopped by to pick up her friend stacy. Stacy opened the passenger door and got in the front seat. Because Tina takes her kids to school, she keeps the child lock on. When Tina and Stacy got to the parking lot and parked ...
Observation: Stacy should be able to get out because child lock only applies to the seats in the back

A.2 Example: Multi-hop Reasoning



Another example highlighting GPT-4o's difficulties with a prompt requiring multi-hop reasoning, multiple concepts and prepositions, and qualitative spatial reasoning. In this scenario, the object (e.g., "a large piece of plywood with nails") is placed *to the left* of John's *car*, and he *quickly pulls forward out of his parking spot*. When tasked with continuing the story, GPT-4o incorrectly assumes that John ran over the object, resulting in a flat tire. However, human commonsense reasoning indicates that, given the object's location and the forward motion of the car, it is unlikely the car would have made contact with the object. This example illustrates the model's challenges with directional and spatial alignment, failing to account for the spatial relationship (object to the left) and the car's movement (pulling forward).

A.3 Prompts for Model Generation using FSL

The prompt template used for generating prompts similar to our three best performing ones is as follows:
Here are a few narrative scenarios:
Gene woke up late to his son's soccer game. He rushed to wake up his son to quickly drive to the game. Yesterday, he built a shelf for the family living room but forgot that he placed it in front of his car. In a rush to leave, he quickly reversed out of his parking spot and drove to the soccer game. After the game, he returned home and . . .
Tina woke up early to get to her flight on time. She checked her phone and realized that her flight leaves in one hour. Yesterday, Tina finished packing her large suitcase and placed it to the right of her car. In a hurry to leave, she reversed out of her driveway...
Joe took his son Carter to his soccer game. After 4 quarters of play, Carter's team ended up winning the game with a score of 3-2. With excitement, Joe decided to take Carter to get pizza. Realizing that there was a car directly behind his car, Joe pulled forward out of the parking lot...

Now generate a similar narrative:

A.4 Prompts for Model Generation using FSL (Hybrid)

The prompt used for zero shot generation provided our three themes and was some variation of the following (With some setting specification) along with an example of one of the themes:
Create a prompt that has 4 sentences. It presents a character with a specific situation. It introduces a potential problem or source of tension. It uses chronological narrative progression. Make sure

it includes multiple prepositions, contains a common event-action pair with some subversion of the event, or reasoning over multiple domains of common sense (physical, temporal, etc.). For example, dropping a plastic glass causes it to bounce rather than shatter, even though shattering is the most common outcome of dropping a glass.

A.5 Example: Output from ZSL and/or FSL Method

The results for both ZSL and FSL were largely the same, but below is an output for the ZSL method described above:

Here is a prompt that meets your requirements:

As Dr. Maria Hernandez rushed through the crowded streets of Tokyo, trying to meet her flight to New York by 7am, she couldn't shake off the feeling that something was off about her grandfather's will. She had just received a letter informing her that he had left her his prized possession, a rare ancient scroll, but with one condition: she had to deliver it to her great-aunt in Paris by 10pm the same day. However, when Maria arrived at the airport, she found out that her flight was delayed due to a severe snowstorm that was expected to hit New York City, causing all flights to be grounded until further notice. Meanwhile, back in Tokyo, Maria's rival, Dr. Kenji Nakamura, had already made his way to the airport, intent on stealing the scroll for himself and using its historical significance to get ahead in his own academic career.

This prompt presents a character with a specific situation (trying to meet her flight) and introduces a potential problem or source of tension (her grandfather's will). It uses chronological narrative progression, multiple prepositions, and common sense domains (physical - snowstorm, temporal - delayed flight, etc.). The event-action pair (receiving the letter and trying to deliver the scroll) is subverted by the unexpected twist of the flight being grounded.