Counter-Hypothesis Generation: Towards Evaluating How LLMs Reason About Alternatives

Marzieh Abdolmaleki, Aaron Maladry, Veronique Hoste, Els Lefever
Department of Translation, Interpreting and Communication (LW22), Ghent University
{marzieh.maleki, aaron.maladry, veronique.hoste, els.lefever}@ugent.be

Abstract

Reasoning about alternatives is a fundamental component of human cognition and argumentation, yet it remains unclear whether large language models (LLMs) can coherently generate and assess them. This paper introduces Counter-Hypothesis Generation (CHG), a novel task for evaluating how LLMs construct plausible hypotheses when contextual information changes. Inspired by open-domain commonsense reasoning, where models infer and compare multiple explanations, CHG bridges commonsense and counterfactual reasoning by requiring models to generate hypotheses that remain logically consistent with modified premises. We present a test set annotated by a human expert and complemented with counter-hypotheses generated by OpenAI-o3 and DeepSeek-r1. Experimental results reveal that even advanced reasoning models exhibit notable limitations in counter-hypothesis generation.

Keywords: Counterfactual Reasoning, Hypothesis Generation, Large Language Models

1. Introduction

In argumentation, evaluating a hypothesis often requires considering alternative hypotheses or rebuttals that challenge an initial claim (Toulmin, 2003). Research on mental models (Johnson-Laird, 2010) suggests that humans reason by constructing and comparing multiple mental representations, enabling them to identify relevant features and draw inferences. This process of reasoning about alternatives is central to human cognition, as it allows people to evaluate multiple possibilities and infer causal relationships. For language models, developing a similar ability is crucial to ensure that their reasoning remains intuitive and interpretable to humans. However, whether current LLMs can generate and assess such alternatives in a coherent and contextually appropriate manner remains an open question.

A *counter-hypothesis* is a hypothesis that states something different from, and often opposite to, another hypothesis. Counterfactual reasoning, in turn, occurs when a person mentally modifies a prior factual event and evaluates the consequences of that change. Counterfactual thoughts can be downward, considering how the situation could have been worse, or upward, considering how it could have been better (Roese, 1997). Both counter-hypothesis and counterfactual reasoning involve contemplating alternatives to reality or to an existing hypothesis. However, while counterfactuals concern whether an outcome is better or worse, counter-hypotheses focus on proposing an alternative hypothesis without such evaluative direction.

Existing work on commonsense (Gordon et al., 2012; Sap et al., 2019) and counterfactual reasoning (Qin et al., 2019; Tandon et al., 2019) has advanced our understanding of how models infer or select among competing hypotheses. However, these tasks typically rely on predefined options or focus on causal consistency, leaving open the question of whether models can generate coherent alternatives when the underlying context changes. To address this gap, we introduce the task of **Counter-Hypothesis Generation (CHG)**. Given a premise–hypothesis pair and a controlled modification to the premise, the goal is to generate a new hypothesis that reflects the contextual change while maintaining logical and linguistic coherence. CHG thus bridges **commonsense reasoning**, which concerns plausibility among everyday explanations, and **counterfactual reasoning**, which explores the consequences of hypothetical alterations. By combining these perspectives, the task provides a new framework for evaluating reasoning flexibility in LLMs, an essential aspect of defeasible inference and adaptive language understanding. The main contributions of this work are as follows: (1) we introduce Counter-Hypothesis Generation, a task for evaluating how LLMs reason about alternatives under contextual changes; (2) we present a human-annotated test set; and (3) we evaluate OpenAI-o3 and DeepSeek-R1, highlighting their limitations in maintaining logical consistency and contextual adaptation.

2. Related Work

Research on reasoning and hypothesis generation in NLP has evolved along three main, complementary lines: commonsense reasoning, counterfactual reasoning, and abductive inference. Each addresses different facets of understanding and generating plausible narratives or explanations. In

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5445–5449
11–16 May 2026.

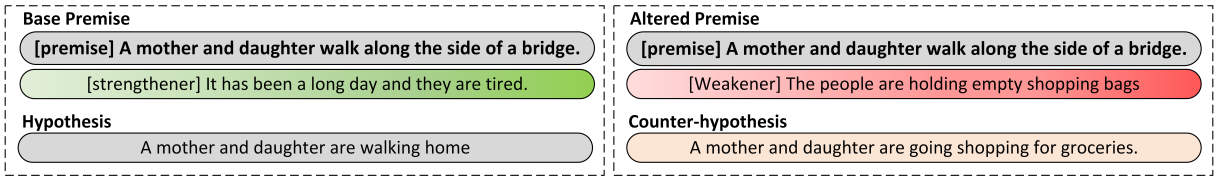

Figure 1: Example from the CHG dataset.

the domain of open-domain commonsense reasoning, benchmarks such as COPA (Gordon et al., 2012), and SocialIQA (Sap et al., 2019) evaluate models' ability to select or predict plausible alternatives for everyday scenarios. These benchmarks primarily test systems on distinguishing between competing explanations or possible continuations. The task in these benchmarks is to predict which of the predefined candidate options is the most probable, often referred to as Natural Language Inference (NLI). By relying on these pre-defined options, these benchmarks cannot evaluate whether generative approaches can produce their own hypotheses. In a related line of research, Abductive Natural Language Inference (αNLI) (Zhao et al., 2023) focuses on identifying or generating missing explanatory hypotheses connecting two observations, but it does not involve modifying an existing premise or hypothesis context. Research on counterfactual reasoning centers around understanding causal relationships under hypothetical changes. Benchmarks such as TimeTravel (Qin et al., 2019) and WIQA (Tandon et al., 2019) ask models to reason about how altering premises affects outcomes, emphasizing factual or causal consistency. These tasks deepen our understanding of causality in language but do not require models to generate alternative hypotheses within NLI frameworks or to handle context modification in a generative manner. In contrast, our proposed CHG task synthesizes these perspectives. It challenges models to incorporate contextual modifications to a base premise and to generate coherent alternative hypotheses consistent with the altered context. This task uniquely tests models' abilities in defeasible reasoning—the capacity to revise conclusions based on changing information—which is a fundamental aspect of human reasoning and decision-making (Johnson-Laird, 2010; Byrne, 2017). By focusing on generative outputs conditioned on dynamic contexts, CHG provides a novel approach for exploring how LLMs adapt explanations to evolving scenarios.

3. Task Definition and Data Set Construction

The CHG task aims to evaluate whether LLMs can flexibly reason about alternative hypotheses when contextual information changes. Formally, the task takes as input a premise–hypothesis pair $\langle p, h \rangle$ and a modified premise p' that represents a controlled change in the original context. The model must then generate a counter-hypothesis h' that remains coherent and logically consistent with the new premise p', while diverging meaningfully from the original hypothesis h. This formulation captures the cognitive process of reasoning about alternatives, where humans revise conclusions in light of new evidence or altered circumstances. Unlike traditional NLI or counterfactual reasoning tasks that assess classification or causal prediction, CHG explicitly requires generative adaptation: the ability to produce a new, context-appropriate hypothesis.

3.1. Data Source

We construct a new test set based on the δ-NLI corpus (Rudinger et al., 2020), which was originally designed to study defeasible reasoning in NLP.

Defeasible inference allows the strength of a hypothesis H, derived from a premise P, to be modified by introducing additional evidence. Two types of additional statements are defined: a weakener, which decreases the plausibility of H, and a strengthener, which increases it.

The δ-NLI dataset consists of three subsets: δ-SNLI, δ-ATOMIC, and δ-Social, each targeting distinct inference types. δ-ATOMIC emphasizes commonsense and event-based reasoning, while δ-Social captures social interactions. δ-SNLI follows the classical NLI format and is thus the most suitable foundation for constructing the CHG corpus.

3.2. Human Annotation Process

Each row in the dataset contains a (1) *Base Premise*, which is a *Premise* paired with a *Strengthener*, and (2) a corresponding *Hypothesis*. To create a counter-hypothesis, the premise is modified by replacing the Strengthener with a *Weakener*, resulting in an (3) *Altered Premise*. The Human annotator then produces a (4) Counter-Hypothesis that remains coherent with the Altered Premise while diverging meaningfully from the original Hypothesis. To create this dataset, each premise is linked with a strengthener and a weakener from the δ-SNLI dataset. By default, each weakener is paired with a randomly selected strengthener, under the assumption that all strengtheners can

be used interchangeably. When there are more weakeners than strengtheners, each strengthener is used at least once, and then the strengtheners are re-used at random.

In total, 822 instances were annotated by a linguist to produce counter-hypotheses. Figure 1 shows example pairs illustrating the correspondence between weakeners, strengtheners, and annotated counter-hypotheses.

To assess the reliability of the human annotations, 100 samples were independently re-annotated by a second linguist. The two sets of annotations were manually compared to evaluate the consistency of reasoning across annotators. Human evaluation showed that 68.6% of the annotations reflected the same reasoning process and conclusion, suggesting that the weakener guided both annotators toward similar interpretations. This pattern further highlights the inherent subjectivity of CHG, where multiple plausible alternatives can be derived from a single premise.

4. LLM Evaluation

To evaluate the ability of current LLMs to perform counter-hypothesis generation, we conducted zero-shot prompting using two state-of-the-art reasoning models: **OpenAI-o3** (OpenAI, 2025) and **DeepSeek-R1** (Guo et al., 2025). This evaluation provides an initial benchmark based on the CHG test set to assess whether LLMs can adapt hypotheses coherently when the underlying premise is altered.

We designed a structured zero-shot prompt that explicitly specifies the relationship between the Premise, Hypothesis, and Altered Premise. The prompt instructs the model to generate a Counter-Hypothesis that is logically supported by the Altered Premise and semantically distinct from the original Hypothesis. The prompt template is shown in Figure 2.

```
You are given a Base Premise, its Hypothesis,
and an Altered Premise, each enclosed in < >.
The Base Premise contains a statement that
strengthens the Hypothesis ([strengthener]),
while the Altered Premise contains
a statement that weakens it ([weakener]).

Generate a Counter-Hypothesis that:
-- Is logically supported by the Altered Premise
as the Hypothesis is supported by the Base Premise.
-- Differs in meaning from the Hypothesis.
-- Does not repeat exactly what is stated
in the [weakener].
-- Is concise: at most 1 sentence and no more
than 20 words.

Base Premise: <>
Hypothesis: <>
Altered Premise: <>
Counter-Hypothesis:
```

Figure 2: Prompt template for counter-hypothesis generation.

4.1. Automatic Evaluation

The counter-hypotheses generated by the models were compared with the human-annotated references using standard n-gram-based metrics. Specifically, we report the precision-oriented **BLEU-4** score (Papineni et al., 2002), which considers n-grams up to $n = 4$, and the recall-oriented **ROUGE-L** score (Lin, 2004). Since these metrics offer a coarse measure of surface-level similarity, they do not fully capture whether the texts convey a similar meaning. To capture the semantic similarities, we extend our set of evaluation metrics to also include **BERTscore** (Zhang et al., 2020) and **SemScore** (Aynetdinov and Akbik, 2024), two widely-used approaches to evaluate the output of generative models that have shown to correlate with human evaluation. Finally, these automatic metrics have potential for large-scale evaluation but cannot fully capture reasoning quality or contextual coherence. Therefore, they serve as a preliminary quantitative reference for comparison with human judgments.

Metric	OpenAI-o3	DeepSeek-R1
BLEU-4	11.6	17.9
ROUGE-L	37.1	40.5
BertScore	0.76	0.78
SEMScore	0.61	0.62
Human Eval. (%)	41.2	52.2

Table 1: Zero-shot performance on the CHG task.

As shown in Table 1, both models achieve relatively low BLEU and moderate ROUGE scores against human references, indicating notable divergence in lexical choice and phrasing. Although DeepSeek-R1 outperforms OpenAI-o3 across all automatic metrics, both models struggle to generate hypotheses that are coherent and logically consistent with modified premises. This suggests that surface-level similarity alone does not reflect reasoning quality. Human evaluation (see Section 4.2) offers a more reliable perspective, with DeepSeek-R1 attaining better performance than OpenAI-o3. While this approach yields valuable insights, it also highlights the need for more robust automatic metrics, since human evaluation is not a sustainable long-term solution.

4.2. Human Evaluation

To gain more fine-grained insights into the ability of current LLMs to perform counter-hypothesis generation, we designed a human evaluation that goes beyond accuracy. Each item was evaluated along seven dimensions, four qualitative (ordinal) and three NLI-based (binary):

- **Difference (1–3)**: measures how clearly the counter-hypothesis diverges from the original hy-

pothesis.

- **Logical Consistency (1–3)**: evaluates whether the counter-hypothesis is logically sound and consistent with the Altered Premise.
- **Fluency (1–3)**: captures grammaticality and naturalness of the counter-hypothesis.
- **Human-Likeness (binary)**: indicates whether the counter-hypothesis reads as human-produced.
- **CH Neutrality (binary)**: checks whether the counter-hypothesis maintains a neutral tone without bias or contradiction.
- **Stronger CH (binary)**: indicates whether the counter-hypothesis is stronger with the weakener within the Altered Premise.

we also considered each of the different constraints presented in the prompt for human evaluation (Table 2). Human evaluation indicates that 42.3% of OpenAI-o3 outputs and 59.7% of DeepSeek-R1 outputs are logically consistent with the altered premise. Regarding difference from the original hypothesis, 77.7% of OpenAI-o3 and 84.3% of DeepSeek-R1 generations diverge appropriately from the initial hypothesis, demonstrating sensitivity to contextual changes. However, 38.8% of OpenAI-o3 and 28.5% of DeepSeek-R1 outputs repeat elements contained within the weakener, suggesting that both models occasionally overfit to lexical cues rather than performing genuine reasoning-based adaptation.

Constraint	OpenAI-o3	DeepSeek-R1
Logically consistent with altered premise	42.3%	59.7%
Different from hypothesis	77.7%	84.3%
Repeats the weakener	38.8%	28.5%

Table 2: Proportion of LLM outputs under different evaluation constraints.

Metric	Worst	Δ1	Mid	Δ2	Best
Difference	OpenAI-o3	+0.09**	DeepSeek-R1	+0.05!	Human
Logical Consistency	OpenAI-o3	+0.23**	DeepSeek-R1	+0.02!	Human
Fluency	OpenAI-o3	+0.20**	DeepSeek-R1	+0.04**	Human
Human-Likeness	OpenAI-o3	+0.39**	DeepSeek-R1	+0.13**	Human

Table 3: Comparative results across models on qualitative dimensions (** = $p < 0.05$, ! = $p \geq 0.05$).

Table 3 compares OpenAI-o3, DeepSeek-R1, and human references across four qualitative dimensions. OpenAI-o3 consistently performs worst, while DeepSeek-R1 and human outputs achieve higher scores. Significant gains appear in Fluency and Human-Likeness, where human annotations remain more natural and expressive. Logical Consistency and Difference also improves notably for DeepSeek-R1, reaching near-human performance. Overall, DeepSeek-R1 shows clear qualitative improvements over OpenAI-o3, though human texts still represent the highest standard.

4.3. Agreement for Human Evaluation

To analyze the subjectivity of the human judgment on sample-level, we measured inter-annotator agreement across all evaluation dimensions. Two annotators independently evaluated 300 counter-hypotheses (100 per source: Human-Gold, OpenAI-o3, DeepSeek-R1). All items were source-blinded to mitigate bias. For the ordinal dimensions, we report Cohen's κ with quadratic weights (κ_w), which accounts for ordered rating distances. For the binary dimensions, we report unweighted Cohen's κ. We provide results aggregated across all 300 items. Confidence intervals (95% CI) are obtained via non-parametric bootstrap resampling.

Dimension	Agreement (%)	κ / κ_w	95% CI
Difference	55.3	0.216	[0.094, 0.337]
Logical Consistency	60.0	0.173	[0.048, 0.286]
Fluency	52.0	0.124	[0.014, 0.240]
Human-Likeness	60.7	0.260	[0.174, 0.346]
CH Neutrality	89.0	0.339	[0.157, 0.507]
Stronger CH	90.3	0.334	[0.152, 0.521]

Table 4: Inter-annotator agreement on human evaluation.

The results in Table 4 indicate slight to fair agreement across qualitative dimensions, with comparatively higher agreement for the binary NLI-based dimensions. Annotators showed the highest agreement on *CH Neutrality* and *Stronger CH*, indicating that these dimensions are more objective than the others. The modest agreement (slight to fair) for these categories reflects the subjectivity in evaluating the quality of generated counter-hypotheses and should be kept in mind when using the benchmark.

5. Conclusion

This paper (1) introduces the novel Counter-Hypothesis Generation task to evaluate how LLMs reason about alternatives and (2) constructs a novel, publicly available[1] benchmark test set. Initial testing on this benchmark, through both automatic and human evaluation, reveals that even strong reasoning models struggle to generate contextually coherent counter-hypotheses when the underlying premise changes. In future work, we plan to extend the current annotation process to construct a larger dataset for the training and evaluation of counter-hypothesis generation models. In addition, we aim to integrate this framework into the setting of defeasible NLI and evaluate the role of counter-hypotheses in generating update statements.

[1] https://github.com/marzieh-abdolmaleki/CHG

Acknowledgments

This work was supported by the Special Research Fund of Ghent University under grant number BOF.BAF.2024.0248.01. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government – department EWI.

References

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity.

Ruth M. J. Byrne. 2017. Counterfactual thought. *Annual Review of Psychology*, 67:135–157.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Dongxiao Guo, Zhihong Shao, Yihua Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*. Accessed: 2025-10-22.

Philip N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250.

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.

OpenAI. 2025. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-10-22.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Neal J. Roese. 1997. Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148.

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.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

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.

Stephen Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14883–14896, Toronto, Canada. Association for Computational Linguistics.