Abductive Plausibility Is Not in the Embedding: Evidence from a Dual-Hypothesis Analysis

Shahriar Nekouei   Yifei Zhang   Echo Canaday
Department of Data Science, University of Colorado Boulder
Department of Computer Science and Linguistics, University of Colorado Boulder
shahriar.nekouei@colorado.edu, yifei.zhang@colorado.edu, echo.canaday@colorado.edu

Abstract

Abductive event reasoning requires identifying the most plausible explanation for an observed event, yet current language models struggle to distinguish between explanations that are lexically similar but abductively distinct. We hypothesize that abductive plausibility occupies a representational dimension that is largely absent from standard semantic embedding spaces. To investigate this, we introduce a dual-hypothesis framework that contrasts a gold explanation, an evidence-derived explanation, and a deliberately inverted explanation. Using RST-guided hypothesis construction, frozen BGE-small embeddings, and several contrastive and ranking objectives, we attempt to recover an "abductive axis" capable of separating these hypotheses. Across all objectives—including triplet loss, margin ranking, InfoNCE, and difference-vector variants—training collapses: embeddings of all hypotheses remain nearly identical (cosine similarity $\approx$ 0.93–0.94), and no projection head succeeds in recovering abductive structure. These negative results provide strong empirical evidence that abductive distinctions are not encoded in frozen semantic representations and likely require joint fine-tuning or architectures explicitly grounded in causal or discourse-level reasoning. Our findings highlight fundamental limitations of current embedding models for abductive inference and motivate future work on representations that encode explanatory adequacy beyond surface semantics.

1 Introduction

The term abduction was coined by Charles Sanders Peirce in his work on the logic of science. He introduced it to denote a form of non-deductive inference that differs from the more familiar inductive type. Peirce is also widely known as the father of pragmatism, a philosophical tradition which holds that ideas must ultimately be evaluated through their practical and empirical consequences; within this framework, abduction plays a central role.

Historically, abduction has been understood in two closely related but distinct senses. In the first sense, it refers to the role of explanatory reasoning in generating hypotheses. In the second, more common in modern literature, it refers to the role of explanatory reasoning in justifying hypotheses. This latter interpretation is often described as *Inference to the Best Explanation*. In practice, distinguishing which sense of abduction applies in a given case is difficult, and the nature of abduction has remained a topic of philosophical debate.

This debate is not the focus of the present paper. Instead, we take both senses of abduction as relevant background, and we adopt a pragmatic stance tailored to our setting. The present work focuses on abductive event reasoning, particularly in the context of artificial intelligence and, more specifically, large language models. The ambiguity and contested nature of abduction in philosophy carries over into AI research: determining what constitutes a "better" explanation is far from straightforward. Consequently, the ability of AI systems to perform abductive reasoning has become an important research question, given the increasing role of such systems in everyday decision-making and inference tasks.

In the context of explaining the likely cause of an event, subjectivity and bias often influence how explanations are formed. This is evident in news reporting, where different outlets such as CNN and Fox News may offer entirely different interpretations or hypotheses about the same event. Because abductive reasoning naturally allows for multiple plausible explanations, its subjective nature presents challenges for computational modeling.

To make abductive reasoning more tractable for artificial intelligence, we restrict our focus to cases where the explanatory relationship is primarily causal rather than interpretive or opinion-based. By

The implementation is available at https://github.com/Orca71/Task-12-Abductive-Event-Reasoning.

narrowing the task to cause–and–effect reasoning, we reduce the influence of bias and subjectivity on hypothesis formation. For example, given the event “A building in Japan has collapsed and is completely destroyed,” a plausible abductive hypothesis would be “A major earthquake shook the entire region.”

In contexts where the causal connection is direct and well understood, forming a hypothesis is relatively straightforward. However, many real-world events do not admit clear causal explanations. For instance, an event such as “The economy crashed in 2008 because of . . . ” allows for multiple competing hypotheses, each reflecting different theoretical perspectives and interpretations. In such settings, abductive reasoning becomes significantly more difficult because determining the best explanation requires navigating subjective judgments and incomplete information.

Although abductive reasoning in natural language processing is an active area of research, it remains a challenging task for current computational models. Foundational discourse frameworks such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) provide reliable methods for analyzing text by breaking sentences into a nucleus–satellite structure and linking elementary discourse units (EDUs) through relations such as *Reason*, *Evidence*, and *Elaboration*. These relations are semantically and lexically valuable, and they help reveal how different parts of a sentence or passage contribute to its overall meaning.

However, when modeling abductive reasoning for events with multiple plausible explanations, semantic similarity alone becomes insufficient. Many explanations may share highly similar wording while differing significantly in their explanatory quality. This presents a major challenge for large language models: among many lexically similar candidates, how can a system determine which explanation best accounts for the event, and how can it avoid subtle drifts in meaning that lead to incorrect abductive conclusions?

The goal of this paper is to tackle precisely this challenge. One of the core challenges in computational abductive event reasoning lies in how current embedding models represent meaning. Most pretrained encoders map semantically similar sentences to nearby points in vector space, compressing fine-grained distinctions that matter for abductive inference. When multiple explanations share overlapping vocabulary or similar surface forms, their embeddings become nearly indistinguishable, even if their causal implications differ. This geometric collapse makes it difficult for downstream models to detect which hypothesis better accounts for an observed event.

Similarly, likelihood-based scoring from large language models tends to favor fluent or common phrasing rather than explanations that genuinely provide the strongest causal account. As a result, both embedding similarity and LLM probability often fail to reflect abductive plausibility, revealing a fundamental misalignment between common representation methods and the requirements of abductive event reasoning.

Although this paper is developed in the context of addressing Task 12 on abductive event reasoning for the SemEval competition, we also propose our own theoretical perspective on how to approach this challenging problem. Specifically, we aim to isolate and model the abductive component of an event explanation rather than relying on representations that merge abductive reasoning with semantic and lexical similarity.

This abstract framework is intended to capture the explanatory aspects of a hypothesis that make it a plausible cause of an event, even when its surface form or vocabulary is nearly identical to alternative hypotheses. We therefore hypothesize that abductive reasoning occupies a representational space that is partially independent of conventional semantic similarity. In other words, two explanations may appear almost identical lexically and lie close together in an embedding space, yet differ substantially in the strength of their causal support for an event.

This suggests that the abductive dimension of a hypothesis is not naturally encoded in pretrained semantic models and must instead be learned or revealed through an additional mechanism. By treating abductive plausibility as a distinct representational axis, we aim to create a framework in which explanations can be compared not only by what they say but by how well they account for the underlying event.

To investigate this abductive space, we adopt a dual-hypothesis formulation that distinguishes between two types of explanations. The first, denoted H_a, represents a gold or reference hypothesis that is known to provide a valid explanatory account of the event. The second, H_b, is constructed from retrieved evidence in the form of document snippets and therefore reflects how an artificial system

The implementation is available at https://github.com/Orca71/Task-12-Abductive-Event-Reasoning.

might infer an explanation from real-world text.

Although H_a and H_b may share substantial lexical or semantic overlap, their abductive plausibility can differ in subtle but important ways. By comparing these two hypotheses within a controlled representational framework, we aim to identify the dimensions along which a model must discriminate in order to determine which explanation best accounts for the observed event.

Building on this dual hypothesis framework, our approach models abductive event reasoning by separating the semantic content of an explanation from its explanatory adequacy. At a high level, our system constructs structured representations of both H_a and H_b, encodes them using a frozen sentence encoder, and evaluates their plausibility through a downstream scoring model designed to capture abductive distinctions that are not reflected in surface similarity.

By grounding hypothesis construction in discourse cues and by treating abductive plausibility as a representational dimension that must be explicitly learned, our method aims to identify the explanation that best accounts for an observed event even when competing hypotheses appear nearly identical lexically or semantically.

The implementation in this study is available at `our github repo`.

Our contributions are as follows:

1. **Abductive Space Framework:** We propose a conceptual framework that treats abductive plausibility as a representational dimension partially independent from semantic similarity, addressing a key limitation of embedding-based approaches.

2. **Empirical Validation on SemEval Task 12:** We instantiate the proposed framework in a modeling pipeline and validate its effectiveness on the SemEval Task 12 dataset, showing that abductive plausibility can be directly modeled and measured within this formulation.

2 Background and Related Work

Abductive reasoning has only recently been studied within NLP, where the goal is to identify the explanation that best accounts for an observed event. The ART dataset introduced by Bhagavatula et al. (2019) was the first large-scale benchmark for this task, framing abduction as choosing the more plausible explanatory sentence in a short narrative. Although effective for probing basic explanatory inference, ART primarily tests narrative coherence rather than real-world, evidence-grounded reasoning.

Later work incorporated structured knowledge such as event graphs (Du et al., 2021) or commonsense resources to improve explanatory selection. However, most existing approaches continue to rely heavily on semantic similarity or typical event patterns, making it difficult for models to distinguish between explanations that are lexically similar but differ in abductive strength. SemEval Task 12 extends this line of work by focusing on abductive event reasoning grounded in retrieved text, offering a setting where explanations may be semantically close yet differ subtly in how well they account for the event.

Rhetorical Structure Theory (RST), introduced by Mann and Thompson (1987), is one of the most widely used frameworks for modeling discourse structure. In RST, relations are defined between two spans of text, typically a *nucleus* and a *satellite*. The nucleus expresses information that is independently interpretable, while the satellite provides supporting or contextual material that is meaningful only in relation to the nucleus.

RST also segments text into *Elementary Discourse Units* (EDUs), which form the basis for constructing interpretable discourse trees that capture explanatory and supportive relations in a document. Several RST relations are especially relevant for abductive or explanatory reasoning. Below we present a few illustrative examples using original sentences:

- **Reason:** The satellite provides the cause or motivation for the nucleus.
[NUC Maria left the meeting early.] [SAT She suddenly remembered an urgent appointment.]

- **Elaboration:** The satellite offers additional detail about the nucleus.
[NUC The valley was difficult to reach.] [SAT The narrow road leading there was covered in ice.]

- **Evidence:** The satellite supplies information that supports the claim in the nucleus.
[NUC The office must still be open.] [SAT The lights on the third floor are still on.]

- **Attribution:** The satellite identifies the source of information or reported speech.

The implementation is available at `https://github.com/Orca71/Task-12-Abductive-Event-Reasoning`.

[SAT Investigators reported] [NUC that the device malfunctioned due to a wiring defect.]

- **List (Multinuclear):** Multiple nuclei are presented without contrast or comparison.
[NUC The package included a user manual;] [NUC a charging cable;] [NUC and a protective carrying case.]

3 Method

Our dataset consists of multiple-choice abductive reasoning instances. Each instance contains a target event and a set of candidate explanations, with one or more labeled as the gold explanation representing the best abductive account of the event. In addition, the dataset provides a collection of retrieved articles related to the topic of the target event. These documents serve as external evidence from which alternative hypotheses may be constructed.

Our method is based on the idea that abductive plausibility can be examined by forming two hypotheses: a gold-aligned hypothesis H_a and an evidence-derived hypothesis H_b. Both hypotheses are constructed through RST-guided segmentation of the event and candidate explanations into Elementary Discourse Units (EDUs), followed by linking these EDUs using appropriate nucleus–satellite relations.

The first hypothesis, H_a, combines the EDU representation of the target event (as the nucleus) with the EDU representation of the gold explanation (as the satellite). This produces a coherent, discourse-structured hypothesis that reflects the human-annotated abductive explanation.

The second hypothesis, H_b, uses the target event as the nucleus but draws its satellite content from one or more EDUs extracted from retrieved articles associated with the event. This yields a synthetic, document-grounded hypothesis that a model might plausibly infer from real-world text.

For evaluation and analysis, we also construct an additional control hypothesis H_w, which is semantically similar to H_a but abductively incorrect. This allows us to examine whether the model can distinguish between explanations that are close in lexical content yet opposite in explanatory adequacy.

Example. To illustrate the hypothesis construction procedure, consider the target event:

> *"ChatGPT quickly reached one million daily active users."*

The dataset provides four candidate explanations, one of which is labeled as the gold abductive explanation. In this example, the gold explanation is:

> *"Millions of people started using ChatGPT immediately after its launch."*

After EDU segmentation and RST linking, the gold-aligned hypothesis H_a is formed by treating the target event as the nucleus and the gold explanation as its satellite:

> *"ChatGPT quickly reached one million daily active users because millions of people started using it immediately after its launch."*

To construct the evidence-derived hypothesis H_b, we use the target event as the nucleus and extract satellite EDUs from retrieved documents related to the topic. For this instance, one such article states:

> *"The sudden hype surrounding ChatGPT has generated widespread interest since its public release last month."*

This yields:

> *"ChatGPT quickly reached one million daily active users. The sudden hype surrounding the chatbot generated widespread interest upon release."*

For contrastive evaluation, we also generate a wrong-but-semantic neighbor hypothesis H_w, which preserves the lexical content of the gold explanation but inverts its abductive implication:

> *"ChatGPT quickly reached one million daily active users, but this did not occur because millions of people used it immediately after its launch."*

By construction, H_a, H_b, and H_w are lexically similar yet differ in explanatory adequacy.

3.1 Geometric Structure of Abductive Space

The dual–hypothesis formulation naturally implies a geometric relationship between the explanations. Although H_a, H_b, and H_w may be lexically similar, they differ in explanatory role, and these differences can be conceptualized as structured positions within an abductive space.

The implementation is available at https://github.com/Orca71/Task-12-Abductive-Event-Reasoning.

The gold hypothesis H_a represents a coherent causal explanation that directly supports the target event. The wrong-but-semantic neighbor H_w reverses or negates that causal relation. Conceptually, this makes H_a and H_w nearly *orthogonal* in abductive space.

The evidence-derived hypothesis H_b occupies an intermediate position. It is constructed from retrieved documents that are topically related but do not necessarily express a direct causal connection. Thus in abductive space we expect:

$$H_a \longrightarrow H_b \longrightarrow H_w,$$

representing decreasing explanatory adequacy.

This abductive ordering is not visible in semantic embedding space, where all three hypotheses tend to cluster due to shared vocabulary and topic. Our challenge is to construct a representation where abductive distinctions determine direction.

3.2 Conceptual Abductive Roles

Although H_a, H_b, and H_w are lexically similar, they play different abductive roles. The gold hypothesis H_a affirms a causal explanation; H_w negates it. The evidence-derived hypothesis H_b contains partial explanatory value but also noise. Conceptually:

$$H_a \longrightarrow H_b \longrightarrow H_w,$$

representing decreasing explanatory strength. Later sections examine how these roles interact with embedding geometry.

3.3 Encoding Hypotheses into Vector Space

To analyze abductive relationships, we encode each hypothesis into a fixed-length vector using a frozen sentence encoder. This encoder transforms each hypothesis into a dense vector in a shared embedding space.

Using a frozen encoder prevents the model from reshaping the space during training. It ensures the embeddings reflect only the semantic and lexical content present in the text.

However, because H_a, H_b, and H_w share nearly identical surface forms, the encoder produces vectors that lie extremely close together. The encoder captures semantic similarity but not abductive distinctions. This creates a controlled environment: embeddings are semantically meaningful but abductively blind.

The goal of the scoring model is to recover abductive distinctions within this space.

3.4 Abductive Geometry in Embedding Space

We move from conceptual abductive roles to their representation in embedding space. Large language model embeddings encode a mixture of semantic, lexical, syntactic, and pragmatic information. When hypotheses are semantically similar, these factors collapse into a tight region.

Our aim is to isolate the abductive component. We use the embeddings of H_a and H_w as oppositional *anchors*:

$$h_a = \mathrm{Enc}(H_a), \quad h_w = \mathrm{Enc}(H_w).$$

Their difference:

$$h_a - h_w$$

defines a direction along which the only major distinction is abductive plausibility: H_a supports the event; H_w contradicts it.

The embedding of H_b usually lies between them when projected onto this direction. Although the encoder is frozen, training a scoring model helps discover a projection onto a latent abductive axis.

3.5 Scoring Model: Recovering the Abductive Axis

Once the hypotheses H_a, H_b, and H_w have been encoded into embeddings h_a, h_b, and h_w, the remaining challenge is to distinguish between them based on abductive plausibility rather than semantic similarity. Because the encoder is frozen, it cannot adjust the embedding space to reflect abductive structure. Instead, all abductive distinctions must be recovered by a downstream model operating on top of the fixed representation.

To accomplish this, we introduce a lightweight two-layer feedforward scoring model that maps each embedding h to a scalar abductive plausibility score. The first linear layer learns a weighted combination of embedding dimensions, highlighting components that correlate with abductive differences. A ReLU activation introduces a nonlinearity that suppresses dimensions that are not informative for the task. The second linear layer then compresses the resulting representation into a single scalar value:

$$s(h) = W_2 \,\mathrm{ReLU}(W_1 h + b_1) + b_2.$$

The implementation is available at https://github.com/Orca71/Task-12-Abductive-Event-Reasoning.

Although simple, this architecture plays a crucial role in recovering abductive structure. Because h_a and h_w are constructed to be semantically similar yet abductively opposite, their embeddings act as fixed *anchors*. Over many training instances, the model learns to assign consistently higher scores to embeddings aligned with the abductive direction implied by $h_a - h_w$, and consistently lower scores to embeddings aligned with the opposite direction.

The evidence-derived embedding h_b typically falls between h_a and h_w in abductive space. The scoring model learns to place h_b accordingly, assigning it a score that reflects the degree to which it aligns with the abductive direction learned during training. In this sense, the model implicitly discovers a projection vector v such that

$$v^\top h_a > v^\top h_b > v^\top h_w,$$

capturing an ordering that semantic similarity alone fails to reveal.

Importantly, the scoring model does not generate new semantic meaning or modify the underlying encoder. Instead, it extracts the latent abductive signal already present—albeit entangled and collapsed—in the frozen embedding space. This learned projection constitutes the *abductive axis*, a one-dimensional direction along which hypotheses can be ranked by the extent to which they provide plausible causal explanations for a given event.

The next section describes how this scoring model is trained using abductive triples and how its predictions are evaluated within the SemEval Task 12 framework.

3.6 Training Objective

The scoring model is trained using abductive triples (H_a, H_b, H_w), where H_a is the gold explanation, H_b is the evidence-derived hypothesis, and H_w is the semantically similar but abductively inverted hypothesis. After encoding these hypotheses into embeddings (h_a, h_b, h_w), the goal is to learn a scoring function $s(h)$ such that

$$s(h_a) > s(h_b) > s(h_w)$$

for as many training instances as possible. This ordering reflects the abductive geometry described earlier: the gold hypothesis should lie highest on the abductive axis, the wrong hypothesis should lie lowest, and the evidence-derived hypothesis should occupy an intermediate position.

To enforce this ordering, we use a margin-based ranking loss that encourages abductively stronger hypotheses to receive higher scores. Specifically, for each triple we define the loss:

$$\mathcal{L} = \begin{aligned} &\max\bigl(0,\, m - (s(h_a) - s(h_b))\bigr) \\ &\quad + \max\bigl(0,\, m - (s(h_b) - s(h_w))\bigr). \end{aligned}$$

where m is a positive margin hyperparameter. The first term encourages the gold explanation to score higher than the evidence-derived hypothesis by at least m, and the second encourages the evidence-derived hypothesis to score higher than the inverted explanation by at least m. This formulation directly operationalizes the abductive ordering without requiring the model to encode absolute likelihoods.

Because the encoder is frozen, gradients flow only through the scoring model parameters. The model therefore cannot reshape the embedding space; instead, it must learn a direction within the existing representation that best reflects abductive distinctions. Over many training examples, this objective aligns the model's scoring vector with the abductive axis implied by the differences between h_a and h_w.

During training, we minimize the ranking loss across all triples using Adam optimization with a small learning rate to ensure stable convergence. The final trained model thus produces a scalar abductive plausibility score for any hypothesis embedding, enabling the system to select the explanation that best accounts for an observed event.

4 Results

We evaluated our abductive-space framework using a frozen BGE-small encoder and several contrastive objectives designed to recover an abductive axis from hypothesis triples (H_a, H_b, H_w). Across all settings, training collapsed: the scoring models failed to separate gold and incorrect explanations beyond random chance, and the losses converged to trivial solutions.

4.1 Geometry of Frozen Embeddings

Empirically, the encoder maps the observation, gold hypothesis, and wrong hypothesis to almost identical points in embedding space. For many instances we observe:

$$\cos(h_a, h_b) \approx 0.94, \qquad \cos(h_a, h_w) \approx 0.93,$$

The implementation is available at https://github.com/Orca71/Task-12-Abductive-Event-Reasoning.

despite H_b and H_w representing abductively opposite or competing explanations. This confirms that the frozen semantic encoder largely ignores subtle causal and counterfactual structure, instead collapsing hypotheses with similar surface form and topic into a tight cluster.

As a consequence, the model receives almost no discriminative signal at the embedding level: abductive differences are not reflected in the base geometry.

4.2 Triplet and Ranking Objectives

We first trained a projection head with a triplet-style loss, treating (h_a, h_b, h_w) as anchor, positive, and negative:

$$\mathcal{L}_{\text{triplet}} = \max\big(0,\ \|h_a - h_b\|^2 - \|h_a - h_w\|^2 + m\big).$$

In practice, the encoder produced nearly identical distances: $\|h_a - h_b\|^2 \approx \|h_a - h_w\|^2$, so the projection head had no consistent gradient direction along which to pull h_b closer and push h_w away. The model quickly converged to a trivial solution in which all projected embeddings collapse to almost the same vector, driving the loss toward zero without learning meaningful abductive structure.

We then replaced the triplet loss with a pairwise margin-ranking objective enforcing the ordering $s(h_a) > s(h_b) > s(h_w)$. This objective exhibited the same failure mode: because all three embeddings are nearly indistinguishable, the scoring model learns to assign nearly identical scores, and the loss again collapses without improving abductive discrimination.

4.3 InfoNCE and Difference-Vector Variants

To provide the model with stronger relative signals, we experimented with InfoNCE-style contrastive objectives that treat (H_a, H_b) as positives and (H_a, H_w) as negatives. We also tried forming difference vectors, e.g.,

$$d_b = h_b - h_a, \qquad d_w = h_w - h_a,$$

with the intuition that subtracting the observation might isolate the hypothesis-specific component.

Neither variant succeeded. Because BGE-small is a normalized semantic encoder and $h_a \approx h_b \approx h_w$, the differences d_b and d_w are both extremely small and nearly indistinguishable. The projection head again converges to mapping all inputs to a constant (or near-constant) vector, minimizing the contrastive loss without learning an abductive dimension.

Lowering the margin, changing temperatures, increasing training epochs, and adding non-linear layers in the projection head did not alter this behavior. Across all objectives, validation performance remained near random choice among candidates, confirming that no usable abductive signal could be extracted from the frozen embedding space.

5 Discussion

The failure of all contrastive objectives to separate H_a, H_b, and H_w reveals a central limitation of our setting: a frozen semantic encoder does not provide a geometry in which abductive distinctions are recoverable by a lightweight head.

BGE-small is trained primarily for semantic similarity: it is explicitly optimized to compress paraphrases, ignore small lexical differences, and group sentences with similar topical content. From the encoder's perspective, a gold hypothesis, a wrong hypothesis, and an evidence-based hypothesis that all talk about the same event are almost indistinguishable. Causal direction, counterfactual structure, and explanatory adequacy are not part of its training signal, and therefore do not appear as separable dimensions in the embedding space.

Our experiments systematically confirmed this:

- Cosine similarities between h_a, h_b, h_w are extremely high and almost identical, leaving no geometric room for abductive separation.
- Triplet and ranking losses quickly converge to trivial solutions where all projected vectors coincide, yielding near-zero loss without improving task performance.
- InfoNCE and difference-vector variants behave similarly, because subtracting nearly identical embeddings produces nearly zero vectors that are still not distinguishable by a projection head.

Conceptually, this shows that abductive reasoning cannot be obtained "for free" by simply placing a contrastive projection layer on top of a frozen semantic encoder. If the base model has never been exposed to abductive constraints, its representation space contains no gradient signal that separates correct from incorrect causal explanations, even when

The implementation is available at https://github.com/Orca71/Task-12-Abductive-Event-Reasoning.

those explanations are constructed in a carefully controlled way.

This negative result is itself informative. It suggests that:

1. Abductive plausibility is not a linear or low-rank perturbation of semantic similarity in current embedding models; and

2. To model abductive event reasoning, one must either jointly fine-tune the encoder on abductive supervision, or use architectures that explicitly incorporate causal, counterfactual, or discourse-level structure beyond static sentence embeddings.

In other words, our findings support the view that abductive reasoning occupies a representational space that is only weakly, if at all, aligned with the geometry of frozen semantic encoders. The abductive axis we set out to recover does not preexist in BGE-small in a way that can be extracted by a shallow head; it likely must be learned as part of the encoder itself.

References

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

Li Du, Xiao Ding, Ting Liu, and Bing Qin. 2021. Learning event graph knowledge for abductive reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5181–5190, Online. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, University of Southern California, Information Sciences Institute, Los Angeles.

The implementation is available at `https://github.com/Orca71/Task-12-Abductive-Event-Reasoning`.