CE-004

大規模言語モデルにおける暗黙の推論生成能力の評価

根岸直生 [1] 坂口慶祐 [1,2] 乾健太郎 [1,2]
[1] 東北大学大学院 [2] 理化学研究所
naoki.negishi.s5@dc.tohoku.ac.jp

概要

暗黙の推論とは与えられた論証テキスト中の主張・前提間の隔たりを埋める推論であり，暗黙の推論を行う能力は日常の，あるいはディベートの論証等への理解に必要な能力である．本研究では近年の大規模言語モデルの急速な発展をうけ，大規模言語モデルによる暗黙の推論の妥当性についてクラウドソーシングを用いて評価を行った．実験の結果，人手によるアノテーションと同等かそれを上回る評価を全体で得たが，主張と前提に明確で形式的な因果関係が無い場合に妥当な推論を行えない傾向が見られた．暗黙の推論に関するベンチマークは設計が難しいが，人間と同等の推論能力を持つ人工知能の実現に向けて極めて重要である．今回の評価に関するデータをコミュニティに提供することで，後続の研究にて新たな発見や知見を得られることを期待する．

1 はじめに

論証を理解するためには主張と前提という基本的な構成要素を特定し，その間の非明示的な推論（**暗黙の推論**）を認識する必要がある [1]．計算機による暗黙の推論の認識が可能になると，近年活発に研究が行われている論証の自動分析や反論 [2] や学生の論理的思考能力の教育促進 [3] などへ応用可能となる．GPT [4] をはじめとした大規模言語モデルの出現などによる，自然言語処理技術の急速な発展を受け，本研究では大規模言語モデルによる暗黙の推論の認識性能の調査を行った．

本研究で取り扱う **暗黙の推論認識タスク** とは「主張と前提から成る **論証テキスト** から，それらのギャップを埋めるような **暗黙の推論過程** を認識するタスク」である．例えば図 1 は「動物園を廃止すべきである」という主張と「非道徳的であるから」という前提の間の暗黙の推論である．この例における暗黙の推論過程は「動物園を廃止することで，動物が人間の娯楽のために飼育されなくなる．また動物を人間の娯楽のために飼育することは，動物の非道徳的な扱いを引き起こす．(したがって動物園を廃止するべきである)」のような因果推論の形式を持ち，非明示的な知識が挿入されたテキストとして表現される．言語モデルを用いる場合には，論証テキストからこのような暗黙の推論過程を生成するタスクとして扱うことができる．

主張:	動物園を廃止するべきである
前提:	動物の非道徳的な扱いを引き起こすから
暗黙の推論過程:	動物園を廃止すると、 動物が人間の娯楽のために飼育されなくなる。 すると動物は非道徳的な扱いをされなくなる。

図 1: 暗黙の推論認識タスクの概略．モデルは主張と前提に一貫した暗黙の推論過程を出力する．

暗黙の推論の認識に際しては，論証の論理構造や背景知識を捉えることが重要である [2]．言語モデルを利用した先行研究には，論理構造に着目した Betz and Richardson [6] や明示的に背景知識を与える Becker ら [7] による研究などが存在する．これらの先行研究では BART [8] や GPT-2 [9] に対して，大規模な教師データを準備して微調整 (fine-tuning) を行うアプローチが取られている．

一方，近年では 1 千億パラメータを超える大きさの汎用的な大規模言語モデルとして GPT-3 [4] が登場し，既存の自然言語処理ベンチマークタスクに対して，zero-shot, few-shot 事例による in-context learning で非常に高い精度が報告されている．特に，前述の暗黙の推論認識タスクで必要とされる論理構造の把握や世界に関する知識についても，大規模言語モデルはそれらの知識を多く有していることが示唆されている [10, 11]．しかしながら暗黙の推論認識タスクについての定量的な評価はまだ行われていない．そこで本研究では暗黙の推論中の因果関係に着目した IRAC データセット [5] を用いて人手による性能調査を行った．

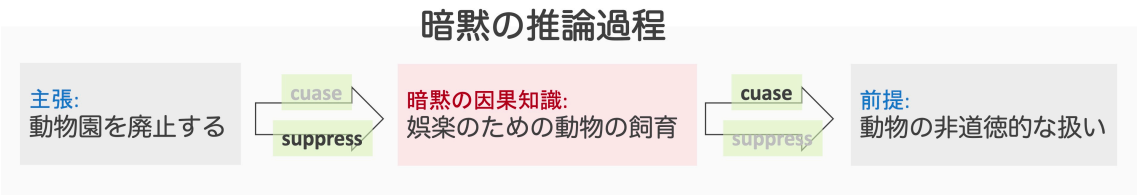

図 2: Singh ら [5] の開発した因果推論フレーム．IRAC (§3) はこのフレームに基づき構築されている．

2 関連研究

古典的論証研究 Toulmin [12] は自然言語による論証の分析に論理学が適していないことを指摘し，今日では Toulmin モデルと呼ばれる論証分析のモデルを構築した古典的な研究である．また論証の形式に関する研究で有名なものに Walton ら [13] による論証スキームがあり，これは 60 個近くの論証における実践的な形式を網羅的に整理したものである．本研究の土台である IRAC データセットもこの論証スキームに着想を得たものである．

Feng and Hirst [14] は Walton らの論証スキームを計算言語学分野へ持ち込んだ代表的な研究の一つであり，論証スキーム中の相当する要素を論証テキストから抽出するタスクを提案した．また Habernal ら [15] はある暗黙の推論とその否定形の推論過程のうち，文脈に照らして妥当なものを選択するという二値分類タスクを提案し，データセットを構築した．

暗黙の推論 Boltužic and Šnajder [16] は暗黙の推論認識タスクに応用できるような大規模なデータセットの構築を初めて試みた研究だが，アノテーション基準にほとんど制約を設けなかったため，推論の深さや内容が多様で扱いにくいアノテーションに終始してしまっている．Becker ら [7] は 5 個程度の前提と単一の主張から成る短い論証に関するデータセットを構築し，鍵となる概念に関する情報を ConceptNet [17] から明示的に背景知識として言語モデルに与えることで暗黙の推論を認識させることを試みた．Chakrabarty ら [18] は NLI データセットと PARA-COMET [19] を用いることで，Becker らと同様に常識的知識に注目した暗黙の推論過程の自動生成を行った．Saha ら [20] は推論過程を明らかにするための説明グラフ ExplaGraphs によってより説明可能性の高い認識方法を提案した．

その中で Singh ら [5] は論証のトピックを区別し，さらに因果関係に限った制約を加えた半構造的なフォーマットで評価の行いやすいデータセットの構築を提案した．

また類似の研究で Betz and Richardson [6] は論証の構造に焦点を当て，再帰的に論証テキストと論理構造をモデルに入力することで論証の再構築を試みる DeepA2 システムを提案した．

3 IRAC データセット

本研究では 2 章の関連研究で紹介したデータセットの中でも Singh ら [5] の IRAC (**I**mplicit **R**easonings in **A**rguments via **C**ausality) を使用する．IRAC は暗黙の推論のうち，実際に行われる割合の高い推論形態である因果推論 [21] [22] に注目したデータセットである．またアノテーションの枠組みも形式的であるように工夫されており，他のデータセットに比べて評価が行いやすい点も特徴である．

IRAC では人工的にデータを収集したコーパス IBM-30K [23] から抽出した論証テキストについて，クラウドソーシングを用いて暗黙の推論過程がアノテートされている．IBM-30K は 71 のトピックを取り扱っているが IRAC はそのうち，死刑制度廃止や動物園廃止等の普遍的なトピック 6 つを利用し，最終的に合計 952 の論証テキストについて 2,636 の暗黙の推論過程のアノテーションを得ている．

3.1 半構造化フォーマット

一般に暗黙の推論過程の形式は自由記述的で非構造的であり，同一の推論に対して無数の記述方法が考えられるため，そのままでは大規模なデータセットを構築するのが難しい．そこで IRAC では図 2 のような因果推論フレームを採用することで，データを半構造化されたフォーマットに制限してアノテーションや評価を行いやすくしている．このフレームは非明示的な推論についての説明に有用であると示されている Walton らの Argument from Consequence Scheme に立脚したものである．

図 1 のように，主張とは論証において表明したい意見を，前提とは主張を支持する根拠を指す．また，それらとは別に図 2 の **暗黙の因果知識** という要素を導入する．これは主張の表現する行動を実行し

た結果であり，かつ前提の表現する事態の原因となるような知識として定義している．Singh らはクラウドソーシングを利用して，因果関係ラベル cause または suppress によって各要素を結んだテキストを，暗黙の推論過程としてアノテートしている．

3.2 データセット品質

Singh らは暗黙の推論過程を収集した後，クラウドソーシングによって論理的妥当性・因果関係の妥当性・キーワードの一貫性の 3 点の品質を保証している．ここでキーワードの一貫性とは，暗黙の因果知識と前提のキーワードとなる単語が類似しているかという点で評価している．

更に専門家 2 人が IRAC から無作為に抽出した 50 例を評価した結果，1 人目は 38 の暗黙の推論過程を，2 人目は 34 の暗黙の推論過程を適格なものとして評価した．ここでの Krippendorff の α 係数 [24] は 0.64 と高い同意度を示したため，データの品質は良好であると結論している．

Singh らによると BART [8] による生成は，トピック特化の微調整を行わない場合で 56 % が妥当と人手評価され，トピック特化の微調整を行った場合では 72 % が妥当と評価された．

4 クラウドソーシング

本研究は先行研究に倣い暗黙の推論認識タスクを暗黙の推論過程の生成タスクとして捉え，入力論証テキストに対する言語モデル出力の人手評価を行った．

まず 4.1 章に示すように各論証テキストについて GPT-3 を用いて暗黙の推論過程を生成した．次に 3 人のクラウドワーカー (以降，評価者と表記) に依頼し，4.2 章で示す基準に従い，文法性や妥当性等の観点から暗黙の推論過程を評価してもらった．その後，集計結果から特に低得点の事例についてさらに著者らが定性分析を行った．また，評価者にはデータセットから抽出した推論過程とモデル出力について，説得性の観点から二者の比較も行ってもらった．

同様の調査は Chakrabarty ら [18] や Becker ら [7] などによって行われているが，単一のデータセットについてより詳細かつ規模の大きな調査を行い定性分析を提供する点で本実験は異なる． また豊富な知識と高い推論能力を有する大規模言語モデルを用いた実験を行う点でも異なる．

なお本実験にて使用したモデルは text-davinci-003 であり，GPT-3.5 と呼ばれるモデルである．[1] また，本実験で取得したデータはすべて GitHub にて公開する．[2]

4.1 推論過程の生成方法

IRAC には 6 つのトピックに関する論証テキストと暗黙の推論のペアが用意されているが，本実験ではそのうち「死刑制度廃止 capital punishment」と「動物園廃止 zoos」という 2 つのトピックに限り調査を行う．この 2 トピックには計 337 の論証テキストと 722 の暗黙の推論が含まれるが，4.3.1 章では全体の約 53 % の 179 の論証テキストを，4.3.2 章では計 60 の論証テキストを取り扱う．これは人手評価を行う都合上，データセットの全例を細かく確認する作業が困難なためである．

本実験では GPT-3 の few-shot learning を利用して各論証テキストに対し推論過程を 1 つ生成し，これをモデルの **生成事例** とした．また IRAC のアノテーションからランダムに抽出した 1 つのゴールド推論過程を **リファレンス事例** としてこれらの評価を行う．なお推論過程が 1 つしかアノテートされていない論証テキストは 4.2 章のリファレンスとの類似性を評価できないため，リファレンス事例としてはデータには含めなかった．

IRAC はテンプレートを用いた構築を行っており文法的なミスが多く含まれていたため，生成や評価にあたり事前に正規表現パターンを利用して二重ピリオドの修正等の最低限の前処理を行った．付録 A に与えたプロンプトや，暗黙の推論過程の詳細な例を示す．

4.2 評価方法

本実験では Becker らの文法性，論理的一貫性 (妥当性)，説明性，リファレンスとの類似性，リファレンスとの比較から説明性を除いた 4 項目 について評価を行った．説明性の評価は主観に依存するため比較評価以外は難しく，また妥当性の基準と評価内容が重複すると判断したためリファレンスとの比較の項目と統合した．

実際には以下に示す項目で絶対評価と相対評価に区別し，Amazon Mechanical Turk を用いて各推論過程につき 3 人の評価者による評価を行った．また，

1） https://platform.openai.com/docs/models/gpt-3-5
2） https://github.com/naoki-negishi/implicit_reasoning

表 1: 4.2.1 章の各項目における暗黙の推論過程の評価結果．文法性と妥当性は 5 段階評価

		文法性	妥当性	類似性
死刑廃止	GPT-3	**3.91**	3.72	0.868
	人間	3.41	**3.73**	-
動物園廃止	GPT-3	**4.01**	3.97	1.10
	人間	4.00	**4.09**	-

絶対評価項目の妥当性において 3 点以下の低得点な事例について，原因を特定するために著者により主張から暗黙の因果知識の説明が妥当でない，暗黙の因果知識から前提への説明が妥当でない，ともに妥当でない，ともに妥当の 4 分類を行った．

なお正確な評価が難しいこと [7] から，本実験では Sentence-BERT 等の自動評価指標は用いなかった．

4.2.1 絶対評価項目

絶対評価において死刑廃止のトピックにて生成事例は 101 事例，リファレンス事例は 64 事例を対象とし，動物園廃止トピックでは生成事例は 78 事例，リファレンス事例は 70 事例を対象とした．

文法性 (1-5 点) 文法，構文的な誤りが存在しなく，更に流暢である場合に 5 点．

妥当性 (1-5 点) IRAC の論証構造は因果関係に立脚したものであった．したがって生成される暗黙の推論過程も因果関係が妥当であることが期待され，そのような場合に 5 点．

リファレンスとの類似性 (個数) IRAC では各論証テキストに対し，暗黙の推論が平均 2.9 個アノテートされている．その中でモデルの生成事例と内容的に類似するものの個数をカウントした．

4.2.2 相対評価項目

相対評価においては各トピック 30 の論証テキストについて，生成事例とリファレンス事例を 1 つずつ用意し評価対象とした．

リファレンスとの比較 IRAC はディベートコーパスであり，特に口語では表面的な説明力も必要とされるためモデルの生成事例とリファレンス事例について説得性の観点のみから比較を行う．よりもっともらしく，多くの説明を与えるかという観点でモデルの生成事例が優れているか，リファレンスが優れているか，同等かという三値の分類を行った．

表 2: 評価者間の同意度 (Krippendorff の α 係数)

		事例数	文法性	妥当性
死刑廃止	GPT-3	101	.104	.259
	人間	64	.157	.284
動物園廃止	GPT-3	78	.197	.0586
	人間	70	.179	.108

4.3 実験結果

4.3.1 絶対評価

表 1 に示すように GPT-3 の生成事例は人間によるリファレンス事例と比較して文法性の面で上回り，妥当性の面でやや低い結果となった．またリファレンスとの類似性において動物園廃止トピックでは 1.0 を超えており，これは生成事例が IRAC にアノテートされた平均 2.9 個の暗黙の推論過程のうち 1 つは平均的に意味内容で類似していると評価されたことを示す．

したがって動物園廃止トピックでは意味内容的に IRAC と近い推論を認識しているが，因果関係の妥当性の認識あるいは生成時点で人間に劣っていることが示唆される．しかし死刑制度廃止トピックでは逆の傾向が見られ，妥当性の認識と生成能力は同等程度だが意味内容としては IRAC と異なる推論を出力する傾向が見られる．

また，暗黙の推論は主観に依存するという Hulpus ら [2] の指摘の通り，評価者間での評価の同意度は生成事例とリファレンス事例でともに低い値を示した (表 2)．よって非専門家による評価でより質の高い結果を得るためには，再帰的な生成により常識的知識まで落とし込んだ推論過程の提示，価値判断的な記述の排除等，主観に依存しない評価の枠組みを用意する必要がある．

定性分析 妥当性が平均 3 点以下と評価された死刑廃止 45 事例，動物園廃止 19 事例について分類を行った結果，暗黙の因果知識から前提への説明が妥当でないと分類された事例が最も多く，それぞれ 25 事例と 10 事例に及んだ．つまりモデルによる生成事例は 3.2 章におけるキーワードの一貫性に関する制約を満たしていないことが示唆される．

4.3.2 相対評価

4.3 章の絶対評価において生成事例はリファレンス事例に妥当性で僅かに劣っていたが，表 3 に示す相対評価の結果から，両トピックにおいて説得力が

あると評価された数は生成事例のほうが多く，逆の傾向があることがわかる．このことから実践的には，生成事例とリファレンス事例に区別できるほどの差がないことが確認できる．

しかしディベートや論証分析等の高い精度が要求されるタスクでは，依然として因果関係の妥当性の向上が望まれる．

表 3: 各トピックで説得的であると評価された割合

	人間	同等	GPT-3
死刑廃止	36 %	22 %	**42** %
動物園廃止	38 %	22 %	**40** %

5 おわりに

本研究では暗黙の推論認識タスクにおける GPT-3 の性能調査を行うため，IRAC データセットを利用し論証テキストに対する暗黙の推論過程を生成し，それらについてクラウドソーシングを用いて妥当性やリファレンスとの比較などの項目で評価を行った．

その結果，人間と同等程度の因果関係の認識能力を確認することができた一方で，定性分析により因果関係を完全に捉えているわけではないことも明らかになった．しかし相対評価によって，少なくとも人間が普段書く程度の論証であれば差が現れないことも確認でき，この先はより高難易度なデータセットが要求される．

本研究で収集したデータはすべて公開し，後続の研究の役に立つことを願う．

参考文献

[1] Robert H. Ennis. Identifying implicit assumptions. **Synthese**, Vol. 51, pp. 61–86, 1982.

[2] Ioana Hulpus, Jonathan Kobbe, Christian Meilicke, Heiner Stuckenschmidt, Maria Becker, Juri Opitz, Vivi Nastase, and Anette Frank. Towards explaining natural language arguments with background knowledge. In **PROFILES/SEMEX@ISWC**, 2019.

[3] Sarah von der Mühlen, Tobias Richter, Sebastian Schmid, and Kirsten Berthold. How to improve argumentation comprehension in university students: experimental test of a training approach. **Instructional Science**, Vol. 47, No. 2, pp. pp. 215–237, 2019.

[4] 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. Language models are few-shot learners, 2020.

[5] Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naito, and Kentaro Inui. IRAC: A domain-specific annotated corpus of implicit reasoning in arguments. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 4674–4683, Marseille, France, June 2022. European Language Resources Association.

[6] Gregor Betz and Kyle Richardson. DeepA2: A modular framework for deep argument analysis with pretrained neural Text2Text language models. In **Proceedings of the 11th Joint Conference on Lexical and Computational Semantics**, pp. 12–27, Seattle, Washington, July 2022. Association for Computational Linguistics.

[7] Maria Becker, Siting Liang, and Anette Frank. Reconstructing implicit knowledge with language models. In **Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures**, pp. 11–24, Online, June 2021. Association for Computational Linguistics.

[8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[9] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[10] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[12] Stephen E. Toulmin. The uses of argument. 1960.

[13] Douglas Walton, Chris Reed, and Fabrizio Macagno. Argumentation schemes. 2008.

[14] Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In **Annual Meeting of the Association for Computational Linguistics**, 2011.

[15] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1930–1940, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] Filip Boltuzic and Jan najder. Fill the gap! analyzing implicit premises between claims from online debates. In

ArgMining@ACL, 2016.
[17] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018.
[18] Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. Implicit premise generation with discourse-aware commonsense knowledge models. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6247–6252, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
[19] Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. Paragraph-level commonsense transformers with recurrent memory. In **AAAI Conference on Artificial Intelligence**, 2020.
[20] Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. Explagraphs: An explanation graph generation task for structured commonsense reasoning. **ArXiv**, Vol. abs/2104.07644, , 2021.
[21] Keshav Singh, Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, and Kentaro Inui. Exploring methodologies for collecting high-quality implicit reasoning in arguments. In **Workshop on Argument Mining**, 2021.
[22] Maria Becker, Katharina Korfhage, and Anette Frank. Implicit knowledge in argumentative texts: An annotated corpus. In **International Conference on Language Resources and Evaluation**, 2019.
[23] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis, 2019.
[24] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

A 付録: 具体例

A.1 与えたプロンプトの例

ここでは紙面の都合上 1-shot で例示するが，予備実験にて出力形式を安定させるためには 3-shot が必要なことが明らかとなったため，実際にはすべてランダムサンプリングを行い 3-shot で与えている．

```
 The following "Argument" is composed of pairs of
claims and premises. Please generate "Implicit
Reasoning" in the form of a causal relationship,
such as "Claim causes (or suppresses) Implicit
Reason, and Implicit Reason causes (or suppresses)
Premise."
 Please ensure that while adhering to this format,
the generated implicit reasoning presents a causal
relationship that is natural and coherent within
the context of the sentence.'

Argument:
 We should introduce compulsory voting, a lack of
 voting leads to a situation were their isn't fair
 representation in parliament.
Implicit Reasonings:
 Introducing compulsory voting causes High voter
 participation. And High voter participation causes
 Fair representation.

Argument: We should abolish zoos, they are cruel
 to animals and wildlife
Implicit Reasoning:
```

A.2 具体的な生成例

生成例 1 (低得点)

次の例は文法性 3/2/4 点，妥当性 1/1/1 点と評価された．ダッシュで文が結ばれており，プロンプトで与えた事例と入力形式がかけ離れているため Argument: に続く論証テキスト中の参照箇所の判断が難しく，出力が安定しなかったものと思われる．リファレンスの例として `Abolishing capital punishment suppresses the government doing the inhumane act of taking someone's life. And the government doing the inhumane act of taking someone's life causes No justification for taking a person's life.` というものがある．

```
Argument:
 We should abolish capital punishment, the death
 penalty is outdated and immoral - there is no
 justification for taking a person's life.
Implicit Reasoning:
 Abolishing capital punishment suppresses taking a
 person's life, and taking a person's life causes
 the death penalty being outdated and immoral.
```

生成例 2 (高得点)

次の例は文法性 5/5/4 点，妥当性 5/5/5 点と評価された．リファレンス例は省略する．

```
Argument:
 We should abolish zoos, zoos do more harm than good
 to the animals that are held in captivity largely
 for entertainment purposes.
Implicit Reasoning:
 Abolishing zoos causes animals to no longer be kept
 in captivity. And animals no longer being kept in
 captivity suppresses them being held for entertainment
 purposes.
```