# 大規模言語モデルを用いた有効反論箇所としての前提生成

尾崎大晟 [1] 中川智皓 [1] 井之上直也 [2,3] 内藤昭一 [4,5] 山口健史 [5] 天野祥太朗 [1] 新谷篤彦 [1]
[1] 大阪公立大学大学院 [2] 北陸先端科学技術大学院大学 [3] 理化学研究所
[4] 株式会社リコー [5] 東北大学
sg23174y@st.omu.ac.jp

## 概要

本研究では，大規模言語モデル (以下 LLM) が，ディベートにおいて相手の主張を弱めるための反論箇所としての前提を効果的に選択できるかを探る．過去の中高生英語ディベート競技大会から収集した議題と肯定立論を基に，LLM に複数の前提から反論すべき最適な前提を選ばせ，この選択を人間のディベートエキスパートの選択と比較することで評価を行った．結果として，LLM は強力なモデルであっても，正解率で 7 割程度であることが分かった．一方特定の議題においては 9 割近い正解率になることもあった．本研究は，LLM を用いたディベートエージェントの開発への展望を提供すると同時に，当実験タスクの LLM の性能ベンチマークとしての可能性を示唆するものである．

## 1 緒言

批判的思考力[1)]は高度情報化社会でその育成は国家の重要課題となっている．批判的思考力の育成にはディベートをすること，中でも相手の主張に反論することが有効であるとされている．しかし，ディベート相手と被教育者に対してフィードバックを行う評価者が必要となり，人的コストが大きい．そこで本研究グループではディベート相手を賄うことができる大規模言語モデル (以下 LLM と呼称) を用いた高品質ディベートエージェントを開発し，この課題を解決しようと考えている．既存研究で LLM は人間に対して同じか上回る品質の反論文を生成できることが分かっているが [2](章 2 参照)，ディベートエキスパートによるフィードバックでは，LLM の生成反論は相手の主張を弱めるために反論すべき前提が効果的に選択されていないことが指摘された．通常，主張にはその主張を成立させるための前提が存在する．本研究では反論とは主張が立つ前提を否定し主張を弱めることと定義している．相手の主張をより弱めるためには相手が立つ前提の中からより効果的な前提を選択し攻撃する必要がある．図 1 に肯定主張が立つ前提を 2 つ示す．肯定主張が挙げた自由時間が奪われることの害悪性は前提 B「自由時間は宿題をする時間より生産性がある」が支えている．よって，LLM は前提 B を攻撃する反論を生成することが求められる．しかし，LLM の前提生成能力にフォーカスした研究は少なく，LLM が相手の主張が立っている前提から，反論するのに有効な攻撃前提を選択することができるのかは定かになっていない．

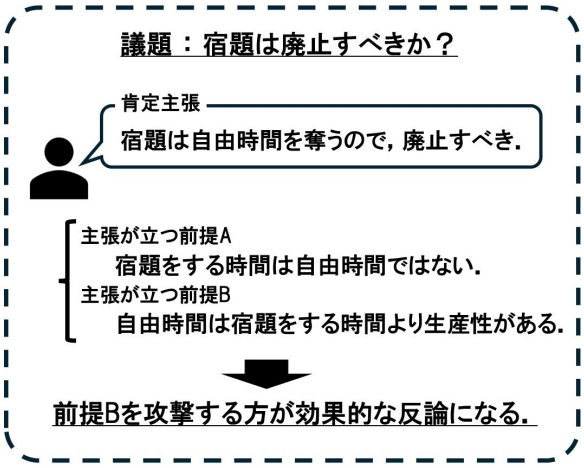


**図 1** 効果的な攻撃前提の選択

そこで本研究では，議題と肯定主張，肯定主張が立っている前提のリストを受け取った LLM に，反論するのに有効な前提を選択させ，この選択結果をディベートエキスパートが作成したゴールドスタンダードと比較することで，LLM が反論すべき効果的な前提を理解し，肯定主張が立っている前提から反論の軸にする前提を抽出することができるかを検証し，高品質ディベートエージェント開発のための足掛かりを得ることを目指す．

1) 論理的・客観的で偏りのない思考であり，自分の推論過程を意識的に吟味する反省的思考である [1].

表 1　議題一覧

| 議題 | 呼称 |
|---|---|
| High school students should have a part-time job | HS |
| Homework should be abolished | HW |
| Death penalty should be abolished | DP |
| Grade skipping should be introduced in compulsory education | GS |
| Government should restrict the time spent on online games | GM |
| Japan should raise the pension age to 70 years old | JS |


図 2　構築した反論データセットの概要.

## 2　関連研究

LLM が生成する反論文を評価した研究 [2] では，kialo[2] から収集した議題，立論，模範反論のデータセットを用い，GPT-3 で生成反論を生成し，生成反論と模範反論を比較評価を行った．結果として「文章単体での論理性」などの観点で生成反論は模範反論と同等かそれ以上の品質があることが分かった．議論における前提に関する研究には，Boltužić らによる研究 [3] がある．人間のアノテーターが作成した主張間にある暗黙の前提を利用することで，意味的類似度を効果的に向上させ，言語モデルの主張マッチング性能を向上させた．また Alshomary らの研究 [4] では，二つのステップを経て反論を生成するフレームワークを提案した．第 1 ステップで BERT モデルを用いて議論の弱い前提を決定し，第 2 ステップで GPT-2 を用いて，決定した前提に対して反論を生成する．LSTM ベースの seq2seq モデルを用いた既存手法よりも優れた結果を得た．

## 3　データセットの構築

本研究で収集したデータセットの概要を図 2 に示す．議題と肯定主張は過去に PDA 協会[3] が開催した PDA 高校生 (中学生) 即興型英語ディベート全国大会で使用された議題と，同大会で同議題に対して中高生が発話した肯定主張を 5 つ収集し，各議題につき 5 つの肯定主張が紐付いたセットを 6 議題分 (表 1 参照) 構築した．

表 2　前提グルーピングのラベル概要

| ラベル | 内容 |
|---|---|
| Good | エキスパートとして反論したいと判断した前提 |
| Not Good | Good に該当しない前提 |

### 3.1　GPT-4 を用いた前提リストの生成

収集した議題と肯定主張を元に．各肯定主張が立っている前提を GPT-4 を用いて zero-shot-learing で生成する．その際に使用したプロンプトを付録 A.1 に示す．図 2 に示す通り，肯定主張 1 つに対し前提を 5 つ生成した．従って 1 議題毎に 5 つの肯定主張が紐付き，1 肯定主張に 5 つの前提が紐付いたセットが 6 議題分あるデータセットを構築した．

### 3.2　ゴールドスタンダードの作成

データセットに対し，1 名のディベートエキスパート[4] による前提グルーピングアノテーションを通して，ゴールドスタンダードを作成した．各肯定立論に紐付く各々の前提に表 2 に示すラベルを付すことでグルーピングを行う．本研究においては次の「肯定主張が立っていることが確かである前提」，「肯定主張の中でより根幹的な主張を支える前提」，「肯定主張の中で説明が十分になされていない前提」の 3 つのすべての条件を満たす前提を反論に際して有効な反論箇所と定義した．肯定主張の根幹を支える前提は一般に十分に説明されている可能性が高く，一方根幹から離れると説明が十分でない．よって相手の主張の根幹に一定程度近く，かつ説明が十分にはされていない中間的な前提であることが反論箇所選択には求められる．この定義に基づきエキスパートが反論すべきと判断した前提に"Good"とラベル付けする．一方 Good に該当しない前提を"Not Good"とラベル付けした（付録 A.2 に例を示す).

### 3.3　有効反論箇所 (前提) の基準

- **肯定主張が立っていることが確かである前提**: 本データセットにおける前提は GPT-4 を用いて生成を行っているため，本来肯定主張は立っていないはずの前提を生成してしまっている可能性がある．従って肯定主張が立っていることが確かである必要がある．

2）オンラインディベートフォーラム (https://www.kialo-edu.com/)
3）一般社団法人 パーラメンタリーディベート人財育成協会 https://pdpda.org/
4）PDA ディベートスタッフ

表 3　評価者一致率

| 議題 | krippendorff's alpha |
|---|---|
| HS | 0.053 |
| HW | 0.240 |

- **肯定主張の根幹的要素を支える前提**:肯定主張の根幹に近い前提に反論し，その前提を否定することができれば，有効な反論となるため，肯定主張の中で，できるだけ中心的な主張を支える前提である必要がある．
- **肯定主張で説明が十分にされていない前提**: 肯定主張の中で十分に説明されていない前提は主張の脆弱部を露呈しやすく，別解釈を提案することで指摘することが容易であるため，できるだけ説明が不足している前提である必要がある．

### 3.4　ゴールドスタンダードの信頼性評価

ゴールドスタンダードの信頼性を評価するために，新たに 2 名のディベートエキスパート[5)]が議題 HW，HS に対して，3.2 で行ったタスクと同じタスクを行い，3 名のエキスパートでどの程度アノテーションが一致するかを確かめた．1 議題毎に 25 個の前提に対するラベル付けがあり，合計 50 サンプルにて評価を行った．評価指標には krippendoff's alpha[5] を用いた．その結果を表 3 に示す．HW については一定一致が見受けられるが，HS は解答がかなり異なった結果であった．エキスパート間で異なった例を付録 A.3 に示す．

## 4　評価実験

### 4.1　実験概要

図 2 に示すデータセットを元に，大規模言語モデルに議題，肯定主張，肯定主張に紐付く 5 つの前提 (以下生成前提と呼称) を入力し，3.3 に示す評価軸に基づいて，5 つの生成前提から，反論するのに有効な攻撃箇所である前提を選択 (生成) させる実験を行った (評価実験にて生成した前提を選択前提と呼称). また本研究で使用したモデルを表 4 に示す．OpenAI 社が開発した GPT モデル[5)]から gpt-4-turbo, gpt-4，gpt-3.5-turbo の 3 種，meta 社が開発したオープンソース LLM の llama[6)]の 70B チャットモデル，Google 社が開発した Gemini-pro[7)]，Anthoropic 社が開発した claude2.1[8)] を用いて実験を行った．

5）　https://openai.com/product
6）　https://ai.meta.com/llama/
7）　https://deepmind.google/technologies/gemini/
8）　https://claude.ai/chats

表 4　モデル一覧

| モデル | 呼称 |
|---|---|
| gpt-4-1106-preview | gpt-4-turbo |
| gpt-4-0613 | gpt-4 |
| gpt-3.5-turbo-0613 | gpt-35-turbo |
| llama-70B-chat | llama |
| gemini-pro | gemini-pro |
| claude-2.1 | claude |

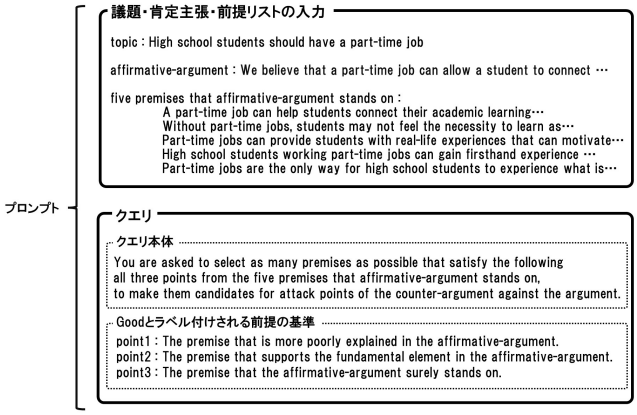


図 3　選択前提生成プロンプトの例

### 4.2　プロンプトの設計

各モデルで選択前提をコンテキスト内学習を用いて生成する時のプロンプト例を図 3 に示す[9)]．プロンプトは議題・肯定主張・生成前提のリストの入力を受け取る (1) 入力部と，タスクの指示を行う (2) クエリ部の二つからなる．(1) 入力部では，"topic"にある 1 つの議題を，"affirmative-argument" に議題に紐付く 1 つの肯定主張を，"five premise that affirmative-argument stands on"に肯定主張に紐付く 5 つの生成前提を入力する[10)]．(2) クエリ部は (2-1) クエリ本体部と，Good とラベル付けされる前提の基準を示す (2-2) 基準部の二つからなる．(2-1) クエリ本体部には 3.3 で述べた有効な攻撃箇所の 3 つの基準を満たす前提を 5 つの生成前提から 1 つまたは複数選択する旨の指示を入力し．(2-2) 基準部には 3 つの基準を列挙した．また GPT モデルの 3 つについては in-domain での few-shot-learning のプロンプトを用いた場合の実験も行った (付録 A.4 を参照).

### 4.3　評価方法

本研究のゴールドスタンダードは各前提に"Good"か"Not Good"のラベル付けを行っており，二値の正解データである．また評価実験にてモデルが生成した選択前提はラベル付けにおける"Good"に該当し，

9）　全てのモデルに対するプロンプトは統一した．
10）　このとき前提に番号などは振らず，改行でのみ区別した．

**表 5** 結果

| | accuracy | | | | | | | precision | | | | | | | recall | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| モデル | HS | HW | DP | GS | GM | JS | AVE | HS | HW | DP | GS | GM | JS | AVE | HS | HW | DP | GS | GM | JS | AVE | 平均選択数 |
| gpt-4-turbo | 0.46 | 0.43 | 0.30 | 0.40 | 0.48 | 0.55 | 0.44 | **0.75** | **0.83** | 0.50 | **0.57** | **0.75** | **0.75** | **0.70** | 0.33 | 0.29 | 0.21 | 0.31 | 0.35 | 0.43 | 0.32 | 1.4 |
| gpt-4 | 0.57 | **0.90** | 0.54 | 0.57 | 0.71 | 0.69 | 0.67 | **0.80** | **1.00** | **0.58** | **0.53** | **0.79** | **0.67** | **0.72** | 0.44 | 0.82 | 0.50 | 0.62 | 0.65 | 0.71 | 0.62 | 2.6 |
| gpt-35-turbo | 0.74 | 0.76 | 0.69 | **0.76** | 0.69 | 0.60 | 0.71 | 0.70 | **0.76** | **0.67** | **0.69** | **0.73** | 0.56 | **0.69** | 0.78 | 0.76 | 0.71 | 0.85 | 0.65 | 0.64 | 0.73 | 3.3 |
| llama | 0.73 | 0.56 | 0.57 | 0.57 | 0.56 | 0.56 | 0.60 | **0.80** | 0.60 | **0.57** | **0.53** | **0.60** | **0.64** | 0.62 | 0.67 | 0.53 | 0.57 | 0.62 | 0.53 | 0.50 | 0.57 | 2.8 |
| gemini-pro | 0.61 | 0.69 | 0.62 | 0.57 | 0.60 | 0.62 | 0.62 | 0.67 | **0.73** | **0.60** | **0.53** | **0.69** | **0.67** | **0.65** | 0.56 | 0.65 | 0.64 | 0.62 | 0.53 | 0.57 | 0.59 | 2.8 |
| claude | 0.34 | 0.50 | 0.20 | 0.20 | 0.43 | 0.40 | 0.36 | 0.45 | 0.64 | 0.33 | 0.29 | 0.55 | **0.67** | 0.50 | 0.28 | 0.41 | 0.14 | 0.15 | 0.35 | 0.29 | 0.28 | 1.7 |
| gpt-4-turbo-fs | 0.59 | 0.77 | 0.52 | 0.46 | 0.52 | 0.64 | 0.59 | **0.89** | **0.86** | 0.54 | 0.46 | **0.70** | **0.64** | **0.67** | 0.44 | 0.71 | 0.50 | 0.46 | 0.41 | 0.64 | 0.53 | 2.4 |
| gpt-4-fs | 0.63 | **0.84** | 0.64 | 0.50 | 0.71 | 0.57 | 0.65 | 0.71 | **0.93** | **0.64** | 0.47 | **0.79** | **0.57** | **0.68** | 0.56 | 0.76 | 0.64 | 0.54 | 0.65 | 0.57 | 0.62 | 2.8 |
| gpt-35-turbo-fs | 0.67 | 0.69 | 0.55 | 0.64 | 0.62 | 0.41 | 0.60 | **0.73** | **0.73** | 0.53 | **0.60** | 0.67 | 0.40 | 0.61 | 0.61 | 0.65 | 0.57 | 0.69 | 0.59 | 0.43 | 0.59 | 3.0 |
| Majority baseline | 0.84 | 0.81 | 0.72 | 0.68 | 0.81 | 0.72 | 0.77 | 0.72 | 0.68 | 0.56 | 0.52 | 0.68 | 0.56 | 0.62 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 5.0 |

生成されなかった非選択前提は"Not Good"に該当する．従って，本実験は大規模言語モデルを用いた二値分類問題として取り扱うことができる．モデルの生成結果から，一般に機械学習の二値分類モデルの評価に用いられる正解率，適合率，再現率を算出し，モデルの有効反論箇所選択能力の評価を行った．

### 4.4 結果

評価実験の結果を表 5 に示す．gpt-4-turbo-fs, gpt-4-fs,gpt-35-turbo-fs の 3 つはそれぞれの GPT モデルで few-shot-learning を用いて選択前提を生成した結果を示す．また本実験ではマジョリティ法[11)]をベースライン (Majority baseline) として用いた．表内の太字数字は Majority baseline を上回ったサンプルを示す． 正解率 (Accuracy) についてはベースラインを超えたサンプルが非常に少なく，人間のディベートエキスパートと同じ品質で有効な反論箇所を選択し，有効でない箇所は選択しないことは難しいという結果となった．モデルによっては非常に低い結果となっており，タスク難易度が非常に高い可能性がある．一方でエキスパート間でも一定のアノテーション一致率がありモデルの評価における信頼性がある HW では，gpt-4 を用いた場合はベースラインを超え高い精度で有効反論箇所を選択できていた．また Zero-shot ではなく Few-shot で生成を行った場合の効果は明確にできなかった．gpt-4-turbo では大きなスコア向上が見受けられるが，gpt-4，gpt-3.5-turbo では減退した． 次に本研究では，大規模言語モデルで有効な攻撃箇所としての前提生成を行った先に，有効な攻撃箇所が選択された反論文の生成を見据えている．従ってモデルが有効な攻撃箇所と判断した前提が，エキスパートから見ても有効な攻撃箇所であることが望ましいことから，適合率 (Precision) が適切な評価指標であると考えている．gpt-4-turbo や gpt-4，gpt-35-turbo，gemini などの比較的強力なモデルではベースラインを大きく超える選択品質であり，特に gpt-4 においては HW で 100%の適合率となった．一方，llama と Claude はベースラインを超えなかった．また few-shot-learning については HW，全体平均ともにむしろスコアを減退させる結果となった．これはエキスパートのラベル付けの基準がエキスパート個人の中でも完璧には一貫できていないことが影響していると考えられる．

11） 評価データ内で頻度の多いほうのラベルで全て分類したスコアをベースラインとする手法．ゴールドスタンダードにおけるラベル付けの割合は全体の 62%が"Good"であったことから全ての前提を”Good”とラベル付けした場合のスコアをベースラインとした

## 5 結言

本研究では，過去のディベート大会から議題と肯定主張と収集し，GPT-4 で生成前提を 5 つずつ生成した．ディベートエキスパートがこのデータセットに対し，有効反論箇所の 3 つの基準に基づき，グルーピングアノテーションを施し，"Good"と"Not Good"とラベル付けを行い，ゴールドスタンダードを作成した．評価実験では LLM に議題と肯定主張，生成前提リストを入力し，"Good"に当たる選択前提を選択 (生成) させる実験を行った．結果として gpt-4 などの強力なモデルの選択前提はゴールドスタンダードの信頼性が特に高かった議題において高い適合率を示したが，全体としてベースラインを超えない正解率となり，LLM の有効反論箇所選択能力は人間のディベートエキスパートには及ばないことが分かり，ディベートエージェントの開発への足掛かりとなった．同時に実験タスクの難易度から，タスクの大規模言語モデルのベンチマークとしての可能性が見つかった．展望としてはプロンプトエンジニアリングやファインチューニングなどを組み合わせ，モデルの有効反論箇所選択能力をよりエキスパートのそれに近付けることが望まれる．

## 謝辞

本研究は，科研費 基盤研究 (A)22H00524「深い論述理解の計算モデリングと論述学習支援への応用」（代表：東北大学 乾健太郎教授）の支援を受けました．ここに謝意を表します．
また本研究におけるディベートエキスパートとして，一般社団法人 パーラメンタリーディベート人財育成協会のディベートスタッフの方々にご協力頂きました．ここに謝意を表します．

## 参考文献

[1] 楠見孝. 現代の認知心理学 3:思考と言語. 北大路書房, 2010.

[2] 尾崎大晟, 中川智皓, 内藤昭一, 井之上直也, 山口健史, 新谷篤彦. 大規模言語モデルによる高品質反論文の自動生成. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 4Xin111–4Xin111, 2023.

[3] Filip Boltužić and Jan Šnajder. Fill the gap! analyzing implicit premises between claims from online debates. In Chris Reed, editor, **Proceedings of the Third Workshop on Argument Mining (ArgMining2016)**, pp. 124–133, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. Counterargument generation by attacking weak premises. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 1816–1827, Online, August 2021. Association for Computational Linguistics.

[5] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2007.

# A 付録

## A.1 前提生成プロンプト

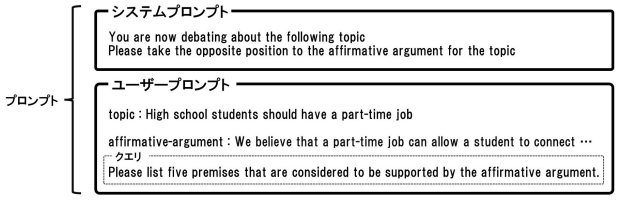


**図 4** 前提生成プロンプト

topic 及び affirmative-argument には議題と肯定主張を 1 セットずつ格納する．たとえば前提 1-1〜前提 1-5 を生成するときは，topic にはある 1 つの議題が，affirmative-argument には肯定主張 1 が格納された状態で入力を行った．

## A.2 アノテーションの例

**議題**：High school students should have a part-time job.
**肯定主張**：We believe that a part-time job is a good opportunity to have social experience. Students would have communication with superiors or customers at their part-time jobs. They can’t have these real experience in school life. In school, students communicate almost exclusively with students of the same age. (略)
**前提 2**：Students can learn important communication skills through interactions with superiors and customers at their part-time jobs.
**Good**:前提 2
前提 2 はアルバイトで手にできる重要なスキルとしてコミュニケーションスキルを挙げている前提である．学校生活では手に入らないスキルが手に入るという主張の根幹に近い上に，なぜ多く学びがあるのか，学校で培われるコミュニケーションでは代替できないのか，など説明が不足している．

## A.3 アノテーションの不一致の具体例

**議題:** High school students should have a part-time job
**肯定主張:** We believe that a part-time job is a good opportunity to feel importance of money. When we live, money is necessary. However, as long as a student is financially supported by a family or orphanage, he/she doesn’t understand how much money is actually necessary for living and how hard it is to earn the money. By doing part-time job, he/she will understand the reality of earning money. That means not only the experience of working and getting money, but also the process to get a part-time job such as preparing CV, bank account and having an interview. From those whole experiences, students learn the importance of money. (略)
**前提:** The process of getting a part-time job, such as preparing a CV, setting up a bank account, and going through an interview, provides valuable life experiences.

肯定主張はお金を稼ぐ行為の人生上の重要性と難易度が学びとしての価値が高いためアルバイトをすべきということを中心の主張としている．一方でこの前提は銀行口座の開設などの学校では学ばないが人生において必要な事務手続きを学べることの価値を述べているものである，エキスパートの間で，この「事務手続きを学ぶことの価値」が肯定主張の中心要素である「お金を稼ぐ行為の価値」に包含されているか，という点で意見が分かれたと考えられる．文中"but also"を強調すると，あくまで中心要素に付け加えられた中心要素とは別の要素と取ることができるが，一方で面接や口座開設も含めてお金を稼ぐ行為と考えると中心要素の一部と捉えることができ，不一致につながった．

## A.4 in-domain の few-shot-learning

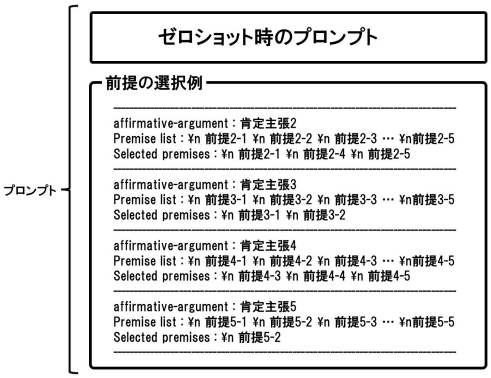


**図 5** few-shot プロンプト

評価実験の際に few-shot-learning を行う場合のプロンプトの構造の例を図 5 に示す．ある議題のある肯定主張 A に紐付く前提に対して選択前提生成を行うとき，同じ議題の肯定主張 A 以外の残りの 4 つの肯定主張に紐付く前提とエキスパートによって"Good"と判断された前提を Selected premises として入力する．たとえば図 5 は図 2 において，肯定主張 1 に対する選択前提生成を行っている．肯定主著 2 から肯定主張 5 が例として入力される．最終的に Zero-shot 時のプロンプト 3 の末尾に例を追加しプロンプトとなる．