主張と根拠の繋がりを評価する科学論文レビュー評価指標

森 清忠 [1,4,*] 田中 翔平 [2] 平澤 寅庄 [2] 小津野 将 [2] 吉野 幸一郎 [3,4,1] 牛久 祥孝 [2],
[1] 奈良先端科学技術大学院大学 [2] オムロン サイニックエックス株式会社
[3] 東京科学大学 [4] 理化学研究所ガーディアンロボットプロジェクト
mori.kiyotada.mh5@naist.ac.jp yoshino.k.ai@m.titech.ac.jp
{shohei.tanaka,tosho.hirasawa,tadashi.kozuno,yoshitaka.ushiku}@sinicx.com

概要

科学論文レビューは，投稿論文数の急増により査読の人的コストの不足に直面している．この問題の解決策として，科学論文レビューの効率化を目的とした言語モデルの活用が積極的に模索されている．人間が理解できる形で，科学論文レビュー評価の裏付け度合いを評価する手法が提案された．その手法では議論の構成要素である主張と根拠を抽出して，主張に根拠がある割合で科学論文レビューの裏付け度合いを評価した．裏付け度合いとは客観的な事実に基づいた主張の度合いである．しかし，科学論文レビューの裏付け度合いの評価では主張に対する根拠の有無だけではなく，主張と根拠の繋がりを適切に評価することでより人間らしい評価を実現できる可能性がある．本研究では主張と根拠の繋がりを評価する新たな科学論文レビュー評価指標を提案し，提案手法と人間の主観評価との相関が従来手法と比べて強いことを示した．

1 はじめに

科学研究の急激な発展により，査読の人的コストの増加が懸念されている．特に，コンピュータ科学のトップ国際会議では投稿される論文数が年々急激に上昇している．しかし，査読者の人数は依然として増加していない [1]．査読者の人員不足により，科学論文レビューの質を確保するための国際会議の主催者の負担は増え続けている [2]．

これらの課題に対処する有望な方法の一つが，言語モデルでの査読支援である．例えば，言語モデルが自動的に低品質なレビューを検出できれば，査読者が自らのレビューを改善する動機付けとなったり，メタレビュアーの意思決定の支援につながることが期待される．この目的の達成のためには，レビューに対して単に品質スコアを付与するだけでなく，人間にとって分かりやすい形で評価を提示することが重要である．

説明可能性 [3] のある科学論文レビュー評価指標として，SubstanScore [4] が提案されている．SubstanScore は，まず科学論文レビューから主観的な意見である主張とその主張を支える客観的な事実である根拠を抽出する．次に，主張に根拠がある割合で科学論文レビューの客観的な事実に基づいた主張の度合いを示す，裏付け度合い [5] を評価する．SubstanScore では最終的な評価値だけではなく，その評価値の算出の根拠となる科学論文レビューに含まれる主張と根拠の注釈が生成される．そのため，人間が解釈可能な形で科学論文レビューの裏付け度合いを評価できる．

SubstanScore は，有名な論証モデルであるトゥールミンの論証モデル [6] の議論の構成要素の内，主張と根拠のみを対象として，科学論文レビューの裏付け度合いを評価した．しかし，科学論文レビューの裏付け度合いは主張と根拠がただあるだけでは十分に評価できない．例えば，主張「私は投稿された論文を不採択とする」に対して，根拠「当該論文の主要なトピックは多くの研究で既に精査されている」は，科学論文レビューに含まれる主張と根拠の組み合わせとして常に正しいとは言えない．これは我々がこの主張と根拠を結ぶ，論証モデルの 3 つ目の主要な要素である論拠，「多くの研究が実施されたトピックは採択されない」をある程度は信じられるが，絶対的に信じられるわけではないと考えるためである．科学論文レビューの裏付け度合いの評価では，主張に対する根拠があるかだけではなく，主張と根拠の繋がりを示す論拠を適切に評価する必要がある．

本研究では，トゥールミンの論証モデルの論拠に

＊ オムロン サイニックエックス株式会社でのインターンシップ期間中に行った研究

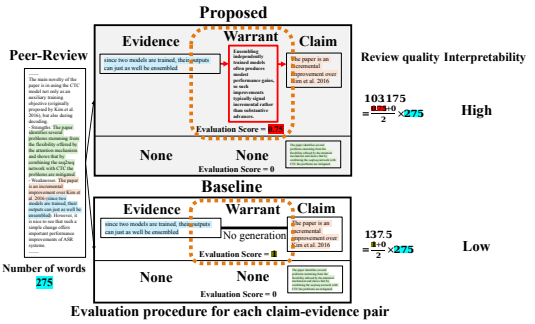

図 1　主張・証拠・論拠を評価する提案された科学論文レビュー評価指標と，主張と証拠のみを評価するベースライン指標との比較

着想を得て，主張に対する根拠の有無だけではなく，主張と根拠を繋ぐ文の妥当性を評価する新たな科学論文レビュー評価指標，WarrantScore を提案する．WarrantScore と SubstanScore の差異を図 1 に示す．WarrantScore は SubstanScore と同様にまずは科学論文レビューに含まれる主張とそれに対する根拠を自動で抽出する．次に，主張と根拠を繋ぐ論拠を生成し評価する．そのため，WarrantScore は SubstanScore よりも説明可能性が高いと考えられる．

我々は WarrantScore と SubstanScore を，人間の裏付け度合いの主観評価が科学論文レビュー文に注釈されたデータセットである SubstanReview [4] と RottenReview [7] で，人間の主観評価との相関の観点から比較評価した．その結果，WarrantScore は SubstanScore よりも人間の主観評価と強く相関することが示された．

2　関連研究

特定の観点を対象とした科学論文レビュー評価指標が提案されている．例えば，厳しさ [8]，徹底性と有用性 [9]，包括性 [5] を対象としている．しかし，これらの科学論文レビュー評価手法は，人間の主観評価値に対する回帰である．そのため，これらの科学論文レビュー評価指標が与えられた科学論文レビューを評価しても，なぜそのように評価したのか説明性がない．我々の提案手法である WarrantScore は，これらの回帰手法と異なり最終的な評価値だけではなく，どのような過程でその評価値が計算されたかを人間が検証できる形で示すことができる．

一般ドメインで文章中に含まれる論拠を大規模言語モデル (LLM) で抽出する試みがある [10]．この研究では，OpenAI が提供する LLM である GPT-4 [11] に，文章の入力の後に，「According to Toulmin model」と追記することで，文章中のトゥールミンの論証モデルに基づく議論の構成要素を自動で抽出した．その結果，GPT-4 が抽出した論拠は 2 名の専門家による評価で 61.7%が受容可能であると示された．論拠の受容可能性は，1) 主張と根拠との関係を完全に説明する十分な関連性があり，2) 自明でなく，3) 注釈者や読者の個人的信念に依らず，根拠から主張を導くために成立すること，の三要件を満たすことと定義されている。

同様に人間が書いた論拠の受容可能性を評価したところ、45.5%の論拠が受容可能であったことから，GPT-4 などの十分なパラメータサイズの LLM は，人間よりも適切な論拠が生成できることが示唆されている．また，この研究では合計で 450 件の論拠と 2 名の専門家による論拠の受容可能性の 2 値評価が含まれたデータセットが公開されている．この研究で公開された知見とリソースは，主張と根拠から LLM で論拠を生成し，その論拠が受容可能かを LLM で分類できることを示唆している．我々の提案手法である WarrantScore では，LLM による論拠の生成と評価の知見を活用し科学論文レビュー評価指標の説明可能性と人間の主観評価との相関を改善することを初めて試みた．

3　提案手法

3.1　SubstanScore

Guo ら [4] は自然言語処理の国際会議に投稿された，550 件の科学論文レビューに対して，主張と根拠を人手で注釈したデータセットを構築した．また，主張に対応する根拠がある割合とレビュー文の単語数を乗算した，科学論文レビューの裏付け度合いの評価指標である SubstanScore を提案した．SubstanScore の定義を式 (1) に示す．

$$\text{SubstanScore} = \frac{\sum_{n=1}^{N} \text{is\_support}(c_n, e_n)}{|C|} \times \text{len(review)} \tag{1}$$

ここで，科学論文レビュー文に含まれる主張の集合を $C = \{c_1, c_2, \cdots, c_N\}$，主張に対する根拠の集合を $E = \{e_1, e_2, \cdots, e_N\}$ と表記する．len(review) は，科学論文レビュー文に含まれる単語数を示し，is_support(c) は e がある場合に 1，ない場合に 0 となる関数である．SubstanScore では，根拠が主張を正当に導くかは検証されていない．

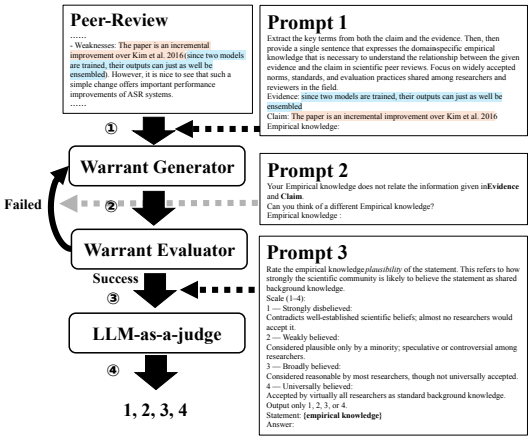

図 2 LLM を用いた論拠の生成と評価のパイプライン：まず，LLM が主張と証拠を繋ぐ論拠を生成する．次に，2 値分類モデルがその受容可能性を 2 値分類する．最後に，別の LLM がその論拠を 4 段階評価する．

3.2 WarrantScore

本研究では，SubstanScore を拡張し，主張と根拠を論理的に結ぶ文である論拠を生成し適切に評価する科学論文レビュー評価指標である，WarrantScore を提案する．WarrantScore の定義を式 (2) に示す．

$$\begin{aligned}\text{WarrantScore} &= \frac{\sum_{n=1}^{N} \mathrm{V}(w_n) \times \text{is\_support}(c_n, e_n)}{|C|} \\ &\quad \times \text{len(review)} \\ &= \text{warrant\_rate} \times \text{len(review)}\end{aligned} \tag{2}$$

ここで，論拠の集合を $W = \{w_1, w_2, \cdots, w_N\}$，論拠 w の評価値を，$\mathrm{V}(w)$ と表記する．WarrantScore では，主張に根拠がある割合だけではなく，主張に根拠がある時の，主張と根拠の繋がりを 0 から 1 の範囲で評価する．WarrantScore を算出するためには，主張と根拠を繋ぐ論拠を生成し，その論拠を適切に評価する必要がある．SubstanScore は WarrantScore における，$\text{is\_support}(c) = 1$ の時に常に $\mathrm{V}(w) = 1$ である場合と解釈できる．

3.3 論拠の生成と評価

WarrantScore の計算のための，主張と根拠からの論拠の生成と生成された論拠の評価のパイプラインを図 2 に示す．本研究での論拠の生成は主張と根拠を結ぶ論拠の受容可能の要件を満たす，査読者の暗黙の常識 [12] を示す文章を生成するものと定義する．論拠の定義は非常に複雑であるため，本研究では科学論文レビュー評価において中心となると考えられる，査読者が過去の査読や研究の経験から獲得したであろう常識のみを対象として生成する．まず，主張と根拠を繋ぐ論拠を LLM で生成する．次に別の LLM で，生成された論拠が受容可能かを 2 値分類し，受容されなかった場合は最大で 3 度，論拠の生成と受容可能かの検証を繰り返す．論拠が 3 回目の分類後も受け入れられない場合には，w がないものとする．

本研究での論拠の評価のタスクは，主張と根拠を繋ぐ文章の事実の妥当性の評価と定義する．LLM-as-a-judge で，その論拠の妥当性を，4 段階リッカート尺度で評価する．WarrantScore の $V(w)$ は，4 段階評価に 0.25 を乗算し 0 から 1 の範囲に変換したものである．SubstanScore と同様に，主張に対する根拠がない場合が $V(w) = 0$ であると考え，等間隔でスコアを割り振っている．また SubstanScore と同様に論拠がない場合は主張と根拠が論理的に繋がっていると解釈し，$V(w) = 1$ にした．実装のための主張と根拠の抽出の詳細は付録 A を参照し，論拠の生成と評価の詳細は付録 B を参照されたい．

4 評価実験

本実験では，WarrantScore が SubstanScore よりも人間の主観評価の裏付け度合いと相関することを確認する．

4.1 評価指標

本実験で評価する，科学論文レビュー評価指標を以下に示す．

- random50%_rate: $\frac{\sum_{n=1}^{N} \text{randint}(1,4) \times 0.25 \times \text{is\_support}(c_n, e_n)}{|C|}$ で示される，全ての論拠へのランダム評価をデータセットに 20,000 回施行した際の人間の主観評価との相関の 50%タイル
- warrant_rate: $\frac{\sum_{n=1}^{N} \mathrm{V}(w_n) \times \text{is\_support}(c_n, e_n)}{|C|}$ で示される，論拠の評価値の平均
- supported_claims: $\frac{\sum_{n=1}^{N} \text{is\_support}(c_n, e_n)}{|C|}$ で示される，主張に根拠がある割合
- Random50%Score: random50%_rate と同様だが $\frac{\sum_{n=1}^{N} \text{randint}(1,4) \times 0.25 \times \text{is\_support}(c_n, e_n}{|C|}$ × review_len に基づく
- WararntScore: warrant_rate × review_len
- SubstanScore: supported_claim × review_len

random50%_rate は提案手法と人間の主観評価との相関の改善が偶然ではないことを示すために採用し

表 1 科学論文レビュー評価指標と人間の主観評価の相関

Metric	SubstanReview	RottenReview
random50%_rate	0.68	0.09
supported_claims	0.45	0.10
warrant_rate	**0.69**	**0.18**
Random50%Score	0.79	0.19
SubstanScore	0.70	0.25
WarrantScore	**0.82**	**0.27**

た．warrant_rate や WararntScore では，論拠の評価のために GPT-5 を用いた．

4.2 データセット

WarrantScore が SubstanScore よりも人間の裏付け度合いの主観評価と相関することを確認するため，科学論文レビューに人間の評価値が注釈されたデータセットである，SubstanReview と RottenReview を用いる．各データセットの概要を以下に示す．

4.2.1 SubstanReview

SubstanReview [4] は，自然言語処理の科学論文レビューデータセットである NLPeer [13] の 550 件の科学論文レビューに主張と根拠を人手でタグ付けしたものである．550 件の内，50 件には 3 人の被験者による裏付け度合いの 3 段階の主観評価が割り振られている．本実験では，人間の裏付け度合いの主観評価が含まれる 50 件のみを，科学論文レビューの評価指標の評価のために用いた．3 人の被験者の裏付け度合いの主観評価は，最大値で統合した．これは科学論文レビューの品質評価という困難なタスクにおける，3 人の被験者間の評価値の分散を抑えるためである．

4.2.2 RottenReview

RottenReview [7] は，*NeurIPs*，*ICLR*，*F1000*，*SWJ* からランダムに抽出された 753 件の科学論文を Objectivity などの複数の観点に基づいて，人間が科学論文レビューをリッカート尺度で 5 段階評価したものである．Objectivity は「Presence of unbiased, evidence-based commentary」を示す．すなわち，SubstanReview の裏付け度合いと類似した評価観点である．異なる被験者が同一の科学論文レビューを評価する場合も確認された．そのため，重複を各被験者の評価値の最大値で統合した．結果として，合計 509 件の科学論文レビュー文が得られた．

表 2 SubstanReview で生成された論拠の具体例

主張	It's great that they also report results on another language,
根拠	showing large improvements over existing work on Japanese CCG parsing.
論拠	When a parsing method produces large gains on Japanese CCG andcomparable gains on another language, it usually signals language-agnostic modeling improvements rather than dataset-specific tricks.

本実験では，Objectivity と科学論文レビュー評価指標との相関で評価指標を評価する．WarrantScore 等の計算のために，SubstanReview で訓練済みの ModernBERT [14] で RottenReview の全ての科学論文レビュー文の主張と根拠を自動タグ付けした．

5 結果

科学論文レビュー評価指標を，評価値と人間の主観評価との相関係数で評価した．SubstanReview と RottenReview での，科学論文レビュー評価指標と人間の主観評価とのスピアマン相関係数を表 1 に示す．SubstanReview と RottenReview では，主張に根拠がある割合のみを考慮した suppored_claim と SubstanScore よりも，それぞれ主張と根拠の繋がりを評価した warrant_rate と WarrantScore が人間の主観評価との相関が高い．これは，科学論文レビューの裏付け度合いを主張と根拠のみで評価するよりも，主張と根拠の繋がりである論拠も評価した方が人間の主観評価と相関することを示している．また，SubstanReviw と RottenReview において，LLM で論拠を評価した場合に，random50%の相関係数を上回った．すなわち，LLM は人間の主観評価との相関を改善するように，適切に論拠を評価できていると考えられる．SubstanReview で LLM が生成した論拠の具体例を表 2 に示す．少なくとも，直感的には，人間が主張と根拠を注釈している SubstanReview では，LLM は与えられた主張と根拠を繋ぐ論拠を生成できていることが示唆されている．

6 おわりに

本研究では，主張に対する根拠の有無だけでなく，主張と根拠を結ぶ論拠を生成し評価する，科学論文レビュー評価指標，WarrantScore を提案した．人間の主観評価が注釈されたデータセットでの評価では，WarrantScore と人間の主観評価との相関は主張と根拠のみを評価した従来手法よりも高いことが示された．

謝辞

本研究は，JST【ムーンショット型研究開発事業】【JPMJMS2236】の支援を受けたものです.

参考文献

[1] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. **arXiv preprint arXiv:2010.05137**, 2020.

[2] Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 4785–4793, 2021.

[3] Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. Towards explainable evaluation metrics for natural language generation. **arXiv preprint arXiv:2203.11131**, 2022.

[4] Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. Automatic analysis of substantiation in scientific peer reviews. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 10198–10216, 2023.

[5] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? **Journal of Artificial Intelligence Research**, Vol. 75, pp. 171–212, 2022.

[6] Stephen E Toulmin. **The uses of argument**. Cambridge university press, 2003.

[7] Sajad Ebrahimi, Soroush Sadeghian, Ali Ghorbanpour, Negar Arabzadeh, Sara Salamat, Muhan Li, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. Rottenreviews: Benchmarking review quality with human and llm-based judgments. In **Proceedings of the 34th ACM International Conference on Information and Knowledge Management**, pp. 5642–5649, 2025.

[8] Rajeev Verma, Rajarshi Roychoudhury, and Tirthankar Ghosal. The lack of theory is painful: Modeling harshness in peer review comments. In **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 925–935, 2022.

[9] Anna Severin, Michaela Strinzel, Matthias Egger, Tiago Barros, Alexander Sokolov, Julia Vilstrup Mouatt, and Stefan Müller. Journal impact factor and peer review thoroughness and helpfulness: A supervised machine learning study. **arXiv preprint arXiv:2207.09821**, 2022.

[10] Ankita Gupta, Ethan Zuckerman, BrendanO'Connor. Harnessing toulmin's theory for zero-shot argument explication. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 10259–10276, 2024.

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

[12] Laurence BonJour. **The structure of empirical knowledge**. Harvard University Press, 1985.

[13] Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. Nlpeer: A unified resource for the computational study of peer review. In **Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)**, pp. 5049–5073, 2023.

[14] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2526–2547, 2025.

[15] Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. A simple and effective model for answering multi-span questions. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 3074–3080, 2020.

[16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. **arXiv preprint arXiv:1606.05250**, 2016.

[17] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In **International Conference on Learning Representations**.

A　主張と根拠の抽出

SubstanScore の実装のため，Guo ら [4] が実施した言語モデルでの主張のタグ付けと主張と対応する根拠の連結の再現実験をし，本実験用に訓練条件を変更した．主張のタグ付けは，与えられたレビュー文から，主張のある複数スパンを抽出するトークン分類タスク [15] である．根拠の連結は，与えられた主張とレビュー文のペアから，対応する根拠のスパンを抽出する質問応答抽出 [16] のタスクである．本研究では，最大入力トークン数が 8,096 トークンである最先端の BERT 系モデル，ModernBERT[14]*で，主張のタグ付けと根拠の連結を再現実験した．本実験の科学論文レビューの主張と根拠のスパンは，SubstanReview の人手で注釈された主張と根拠のスパンから，主張と根拠の両端に存在する，句読点や空白を除去したものである．具体的には「 」 『 』 ( ) ］ } ． ， ！ ？ ； ：と半角スペース，タブ，改行を除去対象とした．SubstanReview の訓練データは 440 件，評価データは 110 件である．訓練データとテストデータを統合した上で 5 分割交差検証を実施し，モデルの性能を評価した．

主張のタグ付けの訓練時のハイパーパラメータは，バッチサイズは 8，エポック数は 20，学習率は 4e-5，weight decay は 0.10，勾配の最大ノルムは 1.0，隠れ層のドロップアウト率は 0.1，注意機構のドロップアウト率は 0.1 である．先行研究では，肯定的な主張と否定的な主張を区別して，BIO（Beginning, Inside, Outside）エンコーディング方式を適用し，B-claim_positive, I-claim_positive, B-claim_negative, I-claim_positive, O の 5 つのクラスのトークン分類を訓練した．しかし，SubstanScore では，肯定的な主張と否定的な主張の区別は考慮されないため，本研究では問題の単純化のため，肯定的な主張と否定的な主張を統合して「主張」としてタグ付けした．そのため，トークン分類のための BIO エンコーディング方式のクラスは B-Claim, I-Claim，O の 3 つである．主張のタグ付けの評価および，後段タスクにおけるスパンは，B-Claim に続く I-Claim が連続し，その後に O が出るまでの範囲と定義した．

根拠の連結の訓練時のハイパーパラメータは，バッチサイズは 4，エポック数は 16，学習率は 1e-5，FP16 による混合精度学習を有効とし，隠れ層のドロップアウト率は 0.2，注意機構のドロップアウト率は 0.2 である．

主張と根拠の抽出を先行研究の評価指標と同じもので，評価した結果を表 3 に示す．これは，訓練済みの ModernBERT が，先行研究の State-of-The-Arts モデルに匹敵する性能を持つことを示す．そのため，本実験での，主張と根拠の抽出では，訓練済みの ModernBERT で主張と根拠を抽出する．最大入力トークン長は 8,096 として，SubstanReview の訓練データ 440 件で学習済みのモデルを用いた．

表 3　関連研究における最先端モデルと我々のモデルの主張と根拠の抽出性能の比較

	Claim tagging			Evidence linking	
	Precision	Recall	F1	EM	F1
Baseline	52.00	59.77	55.61	64.31	82.07
Our	**61.36**	53.13	**56.91**	**66.68**	71.98

B　論拠の生成と評価の実装詳細

論拠の生成には，大規模なオープンソースの言語モデルである GPT-OSS †を，temperature=0 で用いた．論拠の再生成の際は，過去に生成した論拠が，コンテキストとして与えられる．論拠評価の際に，言語モデルが生成する最大出力トークン数は 2 であり，do_sample は false に設定した．

LLM が生成した論拠が受容可能かの推定には，Llama ‡の 2 値分類モデルを用いる．Gupta ら [10] が公開した，450 件の人間の主観による受容評価の付いた論拠を含むデータセットで，Llama を訓練した．Llama の入力を主張，根拠，論拠，目的変数を人間の論拠の受容可能の評価とし，2 値分類のために LoRA [17] で微調整した．その際のデータの分割比率は，訓練データが 7 割，評価データが 3 割である．訓練の際のハイパーパラメータは，エポックは 5，バッチサイズは 4，学習率は 5e-5，入力シーケンスの最大長は 2048 トークン，LoRA のランクは 16，スケーリング係数は 32，ドロップアウト率は 0.05 とした．LoRA を適用するモジュールは q_proj と v_proj である．訓練済みの論拠の受容可能の推定モデルは，評価データセットで，精度が 91.80，再現率が 83.58，F1 スコアが 87.5 である．そのため，訓練済みの論拠の受容可能の推定モデルは，少なくとも一般ドメインの論拠の受容可能を判断できる．

＊ answerdotai/ModernBERT-large

† openai/gpt-oss-120b

‡ meta-llama/Meta-Llama-3-8B