大規模言語モデルを用いた引用文献の重要度分類

大鹿雅史　　笹野遼平
名古屋大学大学院情報学研究科
oshika.masashi.f6@s.mail.nagoya-u.ac.jp
sasano@i.nagoya-u.ac.jp

概要

大規模言語モデルをベースとした文献探索エージェントの発展により，論文執筆に向けた候補文献の検索や推薦が容易になっている．一方で，学術論文における引用文献はその役割や重要度が大きく異なるため，各引用文献を同等に扱うことは適切ではない．既存の引用推薦では主に推薦できたかどうかに着目しているが，重要な論文と重要度の低い論文では引用漏れが執筆論文に与える影響の度合いが異なるため，引用文献の重要度の識別が不可欠である．そこで，本研究では論文内の引用文献の重要度分類に取り組む．具体的には，論文内の引用文脈とメタ情報に基づいて大規模言語モデルが引用文献の重要度を分類可能か検証する．

1　はじめに

学術論文において，論文中で引用される文献はすべてが同等に重要ではない．その論文の貢献を理解する上で不可欠な引用から，一般的な評価尺度や先行研究で利用されたデータセットなど論文の趣旨と関連性の薄い論文の引用まで，その役割と重要度は大きく異なる [1, 2]．論文執筆時には大規模言語モデル（LLM）ベースの文献探索エージェントや引用推薦モデルを利用して候補文献を収集することがあるが，既存の引用推薦の研究ではこれらのモデルを主に正解引用を推薦できたかどうかで評価している．しかし，重要な文献の引用漏れは重要度の低い論文と比較して論文の新規性や貢献の主張に大きな影響を与えるため，引用文献ごとの重要度を識別する必要がある．

引用文献の重要度分類に関する既存研究の多くは，表 1 に示すような引用を含んだ引用文脈や，論文内での言及位置や言及回数などから作成した特徴量をもとに分類を行っている [3, 4, 5]．しかし，これらの手法は引用している論文（引用元論文）の情報は活用しているが，引用されている論文（被引用論文）の情報を多くの場合，活用していないため，二つの論文間の関連性を捉えることができない．

表 1　実際のデータ例．重要度判定の対象となる引用文献は引用文脈の中で#CITATION_TAG に置換される．

引用文脈： One simple but effective strategy is sequential search in which the animal keeps searching until it finds an option that exceeds a threshold of acceptability [2, #CITATION_TAG].
重要度ラベル：1 (重要)
引用文脈： This has found expression in theoretical notions of common ground (Clark and Brennan, 1991), or socially shared cognition (#CITATION_TAG).
重要度ラベル：0 (重要ではない)

そこで，本研究では引用元論文および被引用論文の内容に基づいた引用文献の重要度分類に取り組む．具体的には，引用文脈に加えて，引用元論文および被引用論文のタイトルやアブストラクトといった情報を入力に用いて，その引用が重要か否かを分類する．本研究では，引用元論文と被引用論文の両方の論文の情報を利用した，LLM をベースとした分類モデルを構築し，その有効性を検証する．

2　データセット

論文中の引用文献を対象に，その役割および重要度をアノテーションしたデータセットとして Academic Citation Typing（ACT）[6] が提案されている．ACT は，学術論文の大規模コーパスである CORE [7] で公開されている医学や心理学など複数分野の論文を対象に，論文著者自身が引用文献へのアノテーションを行ったデータセットである．ACT には 883 本の論文から抽出された 11,233 件の引用が含まれ，各引用にはその役割や重要度のラベル，著者情報などが対応づけられている．引用文脈の役割分類や重要度分類を対象としたワークショップの共

通タスク[1)]では，この ACT から 4,000 件の引用文脈を抽出し，3,000 件を訓練セット，1,000 件をテストセットとした ACT SDP [8] が利用されている．本研究では，この ACT SDP に対して被引用論文のタイトルおよびアブストラクト，その他のメタ情報などを付与して拡張された ACT2 [9] を利用する．

表 1 に ACT2 に含まれるデータ例を示す．引用文脈中で分類対象となる引用文献は`#CITATION_TAG`に置換される．この際，引用文脈に複数の引用が含まれる場合でも，評価対象となる 1 件の引用のみが`#CITATION_TAG`に置換される．重要度ラベルは「重要」および「重要ではない」の二値であり，表 2 に示すようにラベル数は均衡している．

3 LLM を利用した分類モデル

本研究では，LLM を利用して引用文脈やタイトルなどの情報をもとに引用の重要度を二値分類するモデルを構築する．出力のスコアに基づく分類境界の調整が可能となるように，モデルは 0 以上，1 以下の連続するスコア s を出力する設定とする．分類境界の調整を行わない場合は，スコアが 0.5 以上の場合は「重要」であり，0.5 未満の場合は「重要ではない」引用として扱う．

具体的なモデル構造として，入力全体の情報を集約した表現として入力の最終トークン位置における隠れ層の埋め込みを線形層に入力して分類を行う．線形層が出力する二つのロジットにソフトマックス関数を適用し，「重要」クラスに対応する値をモデルの出力スコア s として利用する．ファインチューニングは「重要」な引用に対してクラス 1，「重要ではない」引用に対してクラス 0 を正解ラベルとして与え，クロスエントロピー損失を用いて LLM と線形層の重みを更新する．

また，LLM への入力は三種類の設定を用いる．一つ目は引用文脈のみを入力するモデルである（$\mathcal{M}_C$）．二つ目は引用文脈に加え，引用元論文のタイトルとアブストラクトを入力するモデルである（$\mathcal{M}_{C+SP}$）．三つ目は $\mathcal{M}_{C+SP}$ に加えて，被引用論文のタイトルとアブストラクトを入力するモデルである（$\mathcal{M}_{C+SP+CP}$）．これらのモデルを比較することで引用元論文や被引用論文の情報が引用重要度分類の性能に与える影響を検証する．付録 A に本研究で利用したプロンプトを示す．

1） https://www.kaggle.com/c/3c-shared-task-purpose-v2

表 2 利用したデータセットの統計値．

	重要ではない	重要	合計
訓練	1,311	1,189	2,500
検証	257	243	500
テスト	459	541	1,000
合計	2,027	1,973	4,000

4 実験

4.1 実験設定

利用モデル LLM には Qwen3 [10] (8B[2)], 14B[3)]), Gemma-3 [11] (12B[4)]), SciLitLLM1.5 [12] (7B[5)], 14B[6)]) を利用した．SciLitLLM は Qwen2 をベースに学術ドメインに特化させたモデルであり，ドメイン特化モデルの引用重要度分類への有効性を検証するために利用した．また，単一モデルの評価に加え，五つのモデルのモデルスコア s を平均して出力を統合する Ensemble モデルを構築し，複数モデルの出力を考慮する有効性を検証した．

モデルのファインチューニングには QLoRA [13] を利用した．ファインチューニングの設定は最大エポック数を 10 とし，開発セットにおける F 値のマクロ平均が 3 エポック連続で改善しなかった場合に早期終了を行った．学習率は 5×10^{-5}，LoRA の r を 16 とし最適化手法には AdamW [14] を用いた．各モデルについて 10 種類の初期シードで学習を行い，各学習において開発セットにおける F 値のマクロ平均に基づいてモデル選択を行った．最終的な評価として，各シードに対するテストセットでの評価指標の平均および標準偏差を算出した．

データセット 表 2 に本研究で用いたデータセットの訓練・検証・テストセットのデータ数および各分割におけるラベル数を示す．3,000 件の訓練セットを先行研究 [15] に従い，2,500 件の訓練セットと 500 件の検証セットに分割し利用した．

比較手法 比較手法として IREL [15] を用いた．IREL は引用の重要度分類を対象とした 3C Citation Context Classification Shared Task SubtaskB[8)]で 2 位になったモデルであり，SciBERT [16] に引用文脈を入力し，[CLS] トークンの埋め込みをもとに線形層を学習するモデルである．ハイパーパラメータは原論

2） https://huggingface.co/Qwen/Qwen3-8B
3） https://huggingface.co/Qwen/Qwen3-14B
4） https://huggingface.co/google/gemma-3-12b-it
5） https://huggingface.co/Uni-SMART/SciLitLLM1.5-7B
6） https://huggingface.co/Uni-SMART/SciLitLLM1.5-14B
8） https://www.kaggle.com/c/3c-shared-task-influence-v2/

表 3　ACT2 を利用した実験結果．10 個のシード値の平均と標準偏差を示しており，最高スコアを太字で示している．

モデル		マクロ F1	PR-AUC	ROC-AUC
IREL[7)]		0.541 ± 0.020	0.605 ± 0.016	0.573 ± 0.016
Qwen3-8B	$\mathcal{M}_C$	0.537 ± 0.017	0.591 ± 0.014	0.558 ± 0.015
	$\mathcal{M}_{C+SP}$	0.542 ± 0.018	0.598 ± 0.014	0.568 ± 0.023
	$\mathcal{M}_{C+SP+CP}$	0.548 ± 0.019	0.607 ± 0.015	0.571 ± 0.022
Qwen3-14B	$\mathcal{M}_C$	0.551 ± 0.015	0.617 ± 0.020	0.581 ± 0.024
	$\mathcal{M}_{C+SP}$	0.572 ± 0.028	0.623 ± 0.020	0.603 ± 0.022
	$\mathcal{M}_{C+SP+CP}$	0.583 ± 0.016	0.643 ± 0.019	0.616 ± 0.013
Gemma-3-12B	$\mathcal{M}_C$	0.526 ± 0.018	0.598 ± 0.022	0.546 ± 0.015
	$\mathcal{M}_{C+SP}$	0.539 ± 0.019	0.613 ± 0.021	0.577 ± 0.026
	$\mathcal{M}_{C+SP+CP}$	0.554 ± 0.017	0.625 ± 0.024	0.586 ± 0.017
SciLitLLM1.5-7B	$\mathcal{M}_C$	0.535 ± 0.013	0.593 ± 0.016	0.563 ± 0.012
	$\mathcal{M}_{C+SP}$	0.553 ± 0.019	0.621 ± 0.025	0.589 ± 0.024
	$\mathcal{M}_{C+SP+CP}$	0.558 ± 0.015	0.633 ± 0.021	0.592 ± 0.014
SciLitLLM1.5-14B	$\mathcal{M}_C$	0.544 ± 0.013	0.609 ± 0.014	0.575 ± 0.015
	$\mathcal{M}_{C+SP}$	0.545 ± 0.017	0.629 ± 0.019	0.585 ± 0.019
	$\mathcal{M}_{C+SP+CP}$	0.575 ± 0.014	0.641 ± 0.010	0.604 ± 0.016
Ensemble	$\mathcal{M}_C$	0.548 ± 0.014	0.622 ± 0.012	0.581 ± 0.013
	$\mathcal{M}_{C+SP}$	0.570 ± 0.016	0.642 ± 0.014	0.614 ± 0.017
	$\mathcal{M}_{C+SP+CP}$	**0.589** ± 0.009	**0.659** ± 0.010	**0.625** ± 0.008

文で報告されている最適な設定を用い，50 種類のランダムシードで実験を行った．

評価指標　評価指標は，F 値のマクロ平均（マクロ F1），PR 曲線（precision-recall curve）の下部の面積である PR-AUC，ROC 曲線（receiver operating characteristic curve）の下部の面積である ROC-AUC を利用した．F 値のマクロ平均の算出には重要度スコアの閾値を 0.5 とし，PR-AUC および ROC-AUC はこの閾値を変化させることで算出した．

4.2　実験結果

表 3 に実験結果を示す．既存手法との比較において，$\mathcal{M}_C$ は IREL と同等のスコアにとどまっており，言語モデルを近年の大規模なものに置換することによる性能向上は確認できなかった．一方で，利用する論文情報を拡張した $\mathcal{M}_{C+SP}$ と $\mathcal{M}_{C+SP+CP}$ は IREL よりも高い性能を示すことが確認できる．これより，執筆する論文の情報や被引用論文の内容を考慮することが重要度分類の性能向上に寄与することが示された．

モデル間の比較では Qwen3 の 14B モデルがいずれの入力形式においても高い性能を示している．また，SciLitLLM はドメイン特化のモデルであるが，他の汎用モデルと同等のスコアに留まり，明確な性能の向上は確認されなかった．単一のモデルと Ensemble モデルの比較では，すべての入力形式において Ensemble モデルが最高性能を示した．これより，単一のモデルの出力を考慮するだけでなく，複数のモデルの出力スコアを平均することでより高精度な重要度分類が可能であることが示された．

5　分析

本節では，実験結果の分析として，学習データ数および引用方法の違いに着目した分析を行う．

5.1　学習データ数の影響

IREL と比較して LLM を利用した際に分類性能が大きく向上しない理由を分析するために，ファインチューニングに利用するデータ数の影響を調査する．そこで，訓練セットから 25%と 50%に当たる事例を，ラベルの割合を維持したまま無作為に 10 回抽出し，抽出されたそれぞれのデータを利用してモデルのファインチューニングおよび評価を行った．

表 4 にデータ数を変化した際の ROC-AUC の平均および標準偏差を示す．Average は各入力形式ごとに 5 つの LLM および Ensemble モデルのスコアから算出した平均値である．実験結果より，データセッ

8）　共通タスク [8] では IREL のマクロ F1 が約 0.57 と報告されているが，これは複数回提出のうち最高スコアである．本稿でも 50 個のランダムシードで実験した結果，最高値は同程度であった．詳細は付録 B に示す．

表 4　ACT2 からデータをランダムに抽出した際の ROC-AUC の実験結果．10 個のシード値で実験した際の平均と標準偏差を示しており，最も高いスコアを太字で示している．

モデル		25%	50%	100%
Qwen3-8B	$\mathcal{M}_C$	0.537 ± 0.027	0.549 ± 0.022	0.558 ± 0.015
	$\mathcal{M}_{C+SP}$	0.546 ± 0.025	0.555 ± 0.023	0.568 ± 0.023
	$\mathcal{M}_{C+SP+CP}$	0.540 ± 0.018	0.550 ± 0.030	0.571 ± 0.022
Qwen3-14B	$\mathcal{M}_C$	0.546 ± 0.024	0.554 ± 0.019	0.581 ± 0.024
	$\mathcal{M}_{C+SP}$	0.564 ± 0.043	0.574 ± 0.029	0.603 ± 0.022
	$\mathcal{M}_{C+SP+CP}$	0.586 ± 0.034	0.589 ± 0.020	0.616 ± 0.013
Gemma-3-12B	$\mathcal{M}_C$	0.526 ± 0.021	0.551 ± 0.021	0.546 ± 0.015
	$\mathcal{M}_{C+SP}$	0.548 ± 0.026	0.568 ± 0.019	0.577 ± 0.026
	$\mathcal{M}_{C+SP+CP}$	0.555 ± 0.023	0.565 ± 0.024	0.586 ± 0.017
SciLitLLM1.5-7B	$\mathcal{M}_C$	0.555 ± 0.024	0.554 ± 0.023	0.563 ± 0.012
	$\mathcal{M}_{C+SP}$	0.576 ± 0.028	0.572 ± 0.021	0.589 ± 0.024
	$\mathcal{M}_{C+SP+CP}$	0.576 ± 0.022	0.577 ± 0.013	0.592 ± 0.014
SciLitLLM1.5-14B	$\mathcal{M}_C$	0.550 ± 0.027	0.566 ± 0.040	0.575 ± 0.015
	$\mathcal{M}_{C+SP}$	0.558 ± 0.024	0.573 ± 0.021	0.585 ± 0.019
	$\mathcal{M}_{C+SP+CP}$	0.577 ± 0.026	0.594 ± 0.022	0.604 ± 0.016
Ensemble	$\mathcal{M}_C$	0.559 ± 0.022	0.577 ± 0.021	0.581 ± 0.013
	$\mathcal{M}_{C+SP}$	0.581 ± 0.031	0.596 ± 0.018	0.614 ± 0.017
	$\mathcal{M}_{C+SP+CP}$	0.597 ± 0.022	0.608 ± 0.011	**0.625** ± 0.008
Average	$\mathcal{M}_C$	0.546 ± 0.012	0.559 ± 0.011	0.567 ± 0.014
	$\mathcal{M}_{C+SP}$	0.562 ± 0.014	0.573 ± 0.013	0.589 ± 0.017
	$\mathcal{M}_{C+SP+CP}$	0.572 ± 0.021	0.581 ± 0.021	0.599 ± 0.020

トの割合を 25%から 50%，50%から 100%とデータ数を拡張することで，ほとんどのモデルでスコアが向上する傾向が確認できる．ROC-AUC は閾値に依存しない順位付け性能を評価するため，データ増加により重要な引用をより上位に，重要ではない引用をより下位に分離できる方向にモデルが学習したことを示唆しており，学習データ数の増加がさらなる性能の向上につながる可能性を示した[9)]．

5.2　引用方法による影響

次に，引用文脈中における`#CITATION_TAG`の出現形態に着目した分析を行う．表 1 に示すように，ACT2 には`#CITATION_TAG`が単独で出現する事例 (表 1(下)) と，他の引用と並列して出現する事例 (表 1(上)) が存在している．そこで，引用文脈中における`#CITATION_TAG`の出現形態に基づき，評価データを単独引用と並列引用に分割して性能を比較した．テストセット 1,000 件のうち，単独引用が 492 件，並列引用が 508 件であった．

表 5 に Ensemble と Average のマクロ F1 および標準偏差を示す．実験結果より，入力の設定やモデルの種類によらず単独引用が並列引用よりも一貫して高い性能を示していることが確認できる．これは，並列引用では複数文献が同一の文脈を共有するため，`#CITATION_TAG`に固有の手がかりが弱まり，重要度の判断が難しくなる可能性を示唆している．

表 5　引用方法の違いによる性能比較の結果．

モデル		単独引用	並列引用
Ensemble	$\mathcal{M}_C$	0.573 ± 0.018	0.514 ± 0.023
	$\mathcal{M}_{C+SP}$	0.606 ± 0.010	0.526 ± 0.026
	$\mathcal{M}_{C+SP+CP}$	0.625 ± 0.019	0.546 ± 0.017
Average	$\mathcal{M}_C$	0.560 ± 0.007	0.514 ± 0.012
	$\mathcal{M}_{C+SP}$	0.579 ± 0.015	0.522 ± 0.015
	$\mathcal{M}_{C+SP+CP}$	0.599 ± 0.018	0.531 ± 0.014

6　おわりに

本研究では，引用元論文および被引用論文に基づいて LLM が引用の重要度を分類できるか検証した．実験の結果，引用文脈のみを入力とする設定では既存研究と同等のスコアを示した一方で，引用元論文および被引用論文の情報を追加することで性能が向上し，既存手法を上回る結果が得られた．今後は，並列引用など難易度の高い引用形態に対する分類手法の改善とともに，文献収集段階を想定して引用文脈が存在しない状況で重要度を分類可能か検証する．

9)　マクロ F1 および PR-AUC による評価は付録 C に示す．

謝辞

本研究は JST ムーンショット型研究開発事業 JPMJMS2033 および JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（博士後期課程学生支援）JPMJBS2422 の支援を受けたものです.

参考文献

[1] Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. Measuring academic influence: Not all citations are equal. **Journal of the Association for Information Science and Technology (JASIST)**, Vol. 66, No. 2, pp. 408–427, 2015.

[2] Marco Antonio Valenzuela-Escarcega, Vu A. Ha, and Oren Etzioni. Identifying Meaningful Citations. In **AAAI Workshop: Scholarly Big Data**, 2015.

[3] Faiza Qayyum and Muhammad Tanvir Afzal. Identification of important citations by exploiting research articles' metadata and cue-terms from content. **Scientometrics**, Vol. 118, No. 1, pp. 21–43, 2019.

[4] Mingyang Wang, Jiaqi Zhang, Shijia Jiao, Xiangrong Zhang, Na Zhu, and Guangsheng Chen. Important citation identification by exploiting the syntactic and contextual information of citations. **Scientometrics**, Vol. 125, No. 3, pp. 2109–2129, 2020.

[5] Xin An, Xin Sun, and Shuo Xu. Important citations identification with semi-supervised classification model. **Scientometrics**, Vol. 127, No. 11, pp. 6533–6555, 2022.

[6] David Pride, Jozef Harag, and Petr Knoth. Act: an annotation platform for citation typing at scale. In **Proceedings of the 18th Joint Conference on Digital Libraries (JCDL)**, pp. 329–330, 2020.

[7] Petr Knoth and Zdenek Zdrahal. CORE: Three Access Levels to Underpin Open Access. **D-Lib Magazine**, Vol. 18, No. 11/12, 2012.

[8] Suchetha N. Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. Overview of the 2021 SDP 3C Citation Context Classification Shared Task. In **Proceedings of the Second Workshop on Scholarly Document Processing (SDP)**, pp. 150–158, 2021.

[9] Suchetha Nambanoor Kunnath, Valentin Stauber, Ronin Wu, David Pride, Viktor Botev, and Petr Knoth. ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)**, pp. 3398–3406, 2022.

[10] Qwen3 Team. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.

[11] Gemma Team. Gemma 3 technical report. **arXiv preprint arXiv:2503.19786**, 2025.

[12] Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. SciLitLLM: How to Adapt LLMs for Scientific Literature Understanding. In **The Thirteenth International Conference on Learning Representations (ICLR)**, 2025.

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLORA: Efficient Finetuning of Quantized LLMs. In **Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)**, pp. 10088–10115, 2023.

[14] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. **arXiv preprint arXiv:1711.05101**, 2019.

[15] Himanshu Maheshwari, Bhavyajeet Singh, and Vasudeva Varma. SciBERT Sentence Representation for Citation Context Classification. In **Proceedings of the Second Workshop on Scholarly Document Processing (SDP)**, pp. 130–133, 2021.

[16] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific 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)**, pp. 3615–3620, 2019.

表 6　ACT2 からデータをランダムに抽出した際の実験結果．最高スコアを太字で示している．

Model	Input Type	マクロ F1			PR-AUC		
		25%	50%	100%	25%	50%	100%
Qwen3-8B	$\mathcal{M}_C$	0.518 ± 0.039	0.532 ± 0.022	0.537 ± 0.017	0.571 ± 0.032	0.579 ± 0.018	0.591 ± 0.014
	$\mathcal{M}_{C+SP}$	0.519 ± 0.018	0.532 ± 0.019	0.542 ± 0.018	0.590 ± 0.022	0.582 ± 0.019	0.598 ± 0.014
	$\mathcal{M}_{C+SP+CP}$	0.522 ± 0.013	0.531 ± 0.023	0.548 ± 0.019	0.582 ± 0.020	0.592 ± 0.027	0.607 ± 0.015
Qwen3-14B	$\mathcal{M}_C$	0.523 ± 0.021	0.540 ± 0.019	0.551 ± 0.015	0.578 ± 0.023	0.582 ± 0.018	0.617 ± 0.020
	$\mathcal{M}_{C+SP}$	0.536 ± 0.028	0.549 ± 0.027	0.571 ± 0.028	0.598 ± 0.044	0.598 ± 0.024	0.623 ± 0.020
	$\mathcal{M}_{C+SP+CP}$	0.560 ± 0.035	0.564 ± 0.016	0.583 ± 0.016	0.615 ± 0.024	0.616 ± 0.024	0.643 ± 0.018
Gemma-3-12B	$\mathcal{M}_C$	0.515 ± 0.016	0.523 ± 0.016	0.526 ± 0.018	0.574 ± 0.031	0.593 ± 0.018	0.598 ± 0.022
	$\mathcal{M}_{C+SP}$	0.526 ± 0.019	0.540 ± 0.025	0.539 ± 0.019	0.590 ± 0.027	0.601 ± 0.023	0.613 ± 0.021
	$\mathcal{M}_{C+SP+CP}$	0.539 ± 0.022	0.546 ± 0.021	0.554 ± 0.017	0.597 ± 0.023	0.612 ± 0.020	0.625 ± 0.024
SciLitLLM1.5-7B	$\mathcal{M}_C$	0.530 ± 0.016	0.532 ± 0.022	0.534 ± 0.013	0.590 ± 0.029	0.593 ± 0.023	0.593 ± 0.016
	$\mathcal{M}_{C+SP}$	0.548 ± 0.028	0.545 ± 0.017	0.553 ± 0.019	0.611 ± 0.022	0.610 ± 0.016	0.621 ± 0.025
	$\mathcal{M}_{C+SP+CP}$	0.550 ± 0.023	0.553 ± 0.014	0.558 ± 0.015	0.616 ± 0.020	0.621 ± 0.013	0.633 ± 0.021
SciLitLLM1.5-14B	$\mathcal{M}_C$	0.532 ± 0.030	0.541 ± 0.030	0.544 ± 0.013	0.579 ± 0.028	0.597 ± 0.034	0.609 ± 0.014
	$\mathcal{M}_{C+SP}$	0.523 ± 0.013	0.543 ± 0.020	0.545 ± 0.017	0.597 ± 0.026	0.609 ± 0.021	0.629 ± 0.019
	$\mathcal{M}_{C+SP+CP}$	0.556 ± 0.021	0.559 ± 0.017	0.575 ± 0.014	0.609 ± 0.026	0.622 ± 0.024	0.641 ± 0.010
Ensemble	$\mathcal{M}_C$	0.536 ± 0.021	0.551 ± 0.021	0.548 ± 0.014	0.596 ± 0.029	0.608 ± 0.021	0.622 ± 0.012
	$\mathcal{M}_{C+SP}$	0.553 ± 0.026	0.568 ± 0.015	0.570 ± 0.016	0.617 ± 0.032	0.625 ± 0.018	0.642 ± 0.014
	$\mathcal{M}_{C+SP+CP}$	0.568 ± 0.016	0.574 ± 0.008	**0.589** ± 0.009	0.628 ± 0.019	0.643 ± 0.015	**0.659** ± 0.010
Average	$\mathcal{M}_C$	0.526 ± 0.008	0.537 ± 0.010	0.540 ± 0.009	0.581 ± 0.010	0.592 ± 0.010	0.605 ± 0.013
	$\mathcal{M}_{C+SP}$	0.534 ± 0.014	0.546 ± 0.012	0.554 ± 0.014	0.601 ± 0.011	0.604 ± 0.014	0.621 ± 0.015
	$\mathcal{M}_{C+SP+CP}$	0.549 ± 0.017	0.555 ± 0.015	0.568 ± 0.017	0.608 ± 0.016	0.618 ± 0.016	0.635 ± 0.018

表 7　$\mathcal{M}_{C+SP+CP}$ で利用した実際のプロンプト．各プレースホルダーは実際のテキストに置換される．

You are an expert researcher tasked with labeling the importance of a `#CITATION_TAG` in a Citing Context. Output 1 if the `#CITATION_TAG` is central and important to the citing paper's main argument or contribution; otherwise (if it is peripheral or incidental), output 0.
Citing Title:`{citing_title}`
Citing Abstract:`{citing_abstract}`
Cited Title:`{cited_title}`
Cited Abstract:`{cited_abstract}`
Citing Context:`{citation_context}`
Label:

A　プロンプト

表 7 に本研究で利用した $\mathcal{M}_{C+SP+CP}$ のプロンプトを示す．$\mathcal{M}_{C+SP}$ では Cited Title および Cited Abstract を除いたプロンプトを利用し，$\mathcal{M}_C$ では Citing Context のみを利用する．これより，引用周辺文脈に加えて論文情報を段階的に追加した場合の効果を比較する．

B　IREL の詳細

図 1 に IREL を利用して 50 回実験した際のマクロ F1 の分布をヒストグラムで示す．図よりシード値により一定のばらつきが存在することが確認される．また，最小値は 0.486，最大値は 0.576，平均値は 0.541 であり，先行研究で報告されているスコア（約 0.57）は，最大値付近に相当すると考えられる．

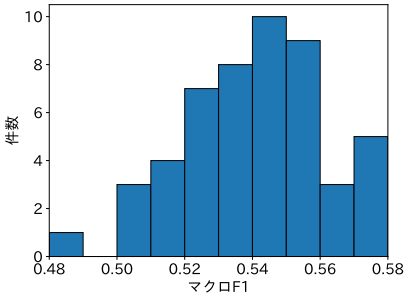

図 1　50 個のシード値で実験した際のマクロ F1 の分布．

C　学習データ数とスコアの詳細

表 6 に 5.1 節と同様にデータ数を変更した際のマクロ F1 および PR-AUC のスコアを示す．実験結果から，ROC-AUC と同様に，PR-AUC は学習データを増加させることで，多くの設定で一貫して改善した．一方で，マクロ F1 はデータセットの割合を 25%から 50%に増やした際には，性能が向上する傾向があるものの，50%から 100%に拡張した際には，$\mathcal{M}_C$ や $\mathcal{M}_{C+SP}$ では同等のスコアに留まった．