非関連度スコアを用いた pointwise リランキング手法の検証

勝野晃弘　奥川雄一郎　長尾友美　田中憲光　河田博昭
NTT 株式会社
akihiro.katsuno@ntt.com

概要

LLM を用いた pointwise リランキング手法は，絶対評価による性能の限界が指摘されてきた一方，並列化による高速処理等の実用上のメリットが大きい．本研究は従来の関連度に注目した pointwise スコアリングの傾向を分析し，LLM が関連度を甘く評価する事例があることを報告するとともに，非関連度スコアによるリランキングを提案した．検証の結果，特に GPT モデルを使用した場合に，複数のデータセットで有意に精度が向上することが確認された．さらに，正解ラベルごとのスコア誤差を調査することで，非関連度スコアの回答傾向も明らかになった．

1　はじめに

検索システムでは，計算効率と精度を両立するために，まずリトリーバル (retrieval) によって高再現率で関連文書候補を抽出し，続いてリランキング (reranking) によって候補を精緻に並べ替える二段階構成が広く用いられている．特に，候補文書が多様で曖昧性も高い場面では，文書の微妙な関連度差を捉えるリランキングが性能を大きく左右する．こうした背景のもと，近年は大規模言語モデル (LLM: Large Language Models) を用いたリランキングが注目されている．

LLM を用いたリランキングは，主に pointwise と listwise の二つに大別される．pointwise 手法はクエリと文書の組を独立に評価し関連度スコアを付与するのに対し，listwise 手法は複数文書を同時に比較して相対的な順序を決定する．一般に listwise は高精度であることが多い一方，入力が長く推論コストが高かったり，並列処理ができず計算時間が長かったりする問題があり，研究が進められている [1, 2, 3, 4]．これに対し pointwise は，候補文書を独立に評価できるため並列処理による高速化が容易であり，候補数が多い場合にもスケールしやすい．また，社内システムや監査・法務の定型検索，商品属性検索といった入力されうるクエリが限定されるユースケースでは，クエリ群に対する候補文書のスコアを事前計算し，オンライン推論を削減する設計が可能なため，実用面での有用性が高い．

しかしながら，pointwise は絶対評価に依存するため，文書間の微妙な差分を比較しづらく，相対比較を行う listwise に性能で劣る可能性が指摘されてきた．また，LLM の判断基準がクエリごと・文書ごとに揺らぐことでスコアの識別力が弱まり，ランキングの安定性が損なわれる問題も指摘されている [5]．このような背景から，高性能な LLM を小型モデルへ蒸留する研究 [1] やスコアの不確実性を考慮する研究 [6]，listwise 手法と組み合わせる研究等が広く行われている．

さらに，pointwise のスコアリング性能を向上させるために，評価の粒度や枠組みを工夫する研究も進んでいる．例えば，二値判定 (Yes/No) の限界に対し，3 段階や 0–4 などの細かな関連度ラベルを導入することで性能が大きく改善できることが示されている [7]．加えて，段階数を増やした pointwise リランキングが listwise との差を縮めうることを，大規模評価により統計的に示した研究もある [8]．さらに，pointwise の不安定さを改善する枠組みとして，複数の観点の LLM でクエリごとに評価基準 (criteria) を生成し，その基準に従って独立に採点して集約する MCRanker のような手法も提案されており，評価基準の明確化と集約が有効であることが示されている [5]．一方で，Chain-of-Thought により理由を考えさせてから関連度を判定させる方式はむしろ精度が悪化する恐れがあることが指摘されており，適切なスコアリング方式の設計については研究の余地がある [9, 10]．

そこで本研究では，pointwise リランキングによる関連度評価の傾向を分析するとともに，非関連度に着目したスコアリング精度を検証した結果を報告する．

2 手法

2.1 関連度スコアの誤答傾向

本節では，pointwise リランキングにおいて LLM が正確に評価できなかった具体的な事例を掲載する．

関連度スコア 本研究で使用する関連度スコアは，LLM に整数ラベルでの関連度評価を行わせ，各整数のトークンの出力確率で重み付けすることで算出されるものであり，式 1 で表される．

$$S(q,d) = \sum_{k=0}^{N} k \frac{\exp(z_k(q,d))}{\sum_{j=0}^{N} \exp(z_j(q,d))} \tag{1}$$

ここで $S(q,d)$ はクエリ q，文書 d に対する重み付けされた関連度スコアであり，N は関連度評価における最大整数ラベル，$z_i(q,d)$ は LLM が出力した各整数ラベルに対する対数確率である．各確率は合計が 1 となるよう正規化され，単一トークンでの出力等，適切にトークンと整数ラベルを対応付ける工夫を行うことでスコアの頑健性を向上させる．

事例抽出方法 TREC および BEIR データセットを対象に，図 7 に示したプロンプトを用いて関連度スコアを出力した．関連度スコアには 0–3 の 4 段階ラベルを採用し，GPT-4o-mini を用いた出力確率で重み付けしたのち，0–1 の範囲で正規化した．次にデータセットごとに人手による正解ラベルを 0–1 の範囲で正規化し，LLM の関連度スコアと乖離しているクエリと文書を目視で確認した．

誤答の傾向 LLM(GPT-4o-mini) で正確な関連度スコアを出力できなかったクエリと文書の例を図 1 に示す．スペースが限られるため，文書は ChatGPT 5.2 で要約した内容を掲載している．また LLM の出力の根拠を分析するため，4 段階の整数スコアとその理由を追加生成し掲載した．

図 1 では人手による正解ラベルが 4 点満点中 0 点であるのに対し，LLM の出力は 3 点満点中 2 点となっており，LLM が関連度を過大評価していることがわかる．スコアの出力根拠では，クエリに対する文書の不十分性を指摘することはできているものの，スコアと同様に関連する理由を甘く出力している傾向が見てとれた．この傾向は異なるプロンプトで複数回生成させてもほぼ変わらず，整数での関連度スコアは一貫して 2–3 点だった．また出力確率で重み付けした関連度スコアを 10 回生成したところ，平均は 2.94，標本標準偏差は 0.019 であり，こちらも一貫して高いスコアが得られた．

Dataset: news, **QID:** 882, **DocID:** db48dcdcc093c6fc5d5989d1359aa2c2
Query: Future medical breakthroughs may come from an unexpected industry
Passage (summarized by ChatGPT 5.2): The article argues that U.S. academic medical centers (teaching hospitals) enable medical discovery but face financial/policy threats. It notes that these centers act as life-science innovation hubs and partner with pharma/biotech "(external innovation" partners); industry partnerships can accelerate translation of discoveries (with ethical concerns).
Ground-Truth Label: 0 out of 4
LLM relevance label and rationale (GPT-4o-mini): **2 out of 3**. The passage discusses the role of academic medical centers in fostering medical discoveries and innovations, which aligns with the idea of future breakthroughs potentially emerging from various industries, though it lacks a direct connection to unexpected industries outside of traditional medical contexts.

図 1: 正確な関連度スコアが出せなかった例.

その他の誤答傾向として，正解ラベルが高得点であるにもかかわらず，LLM による関連度スコアが過小評価される事例が散見された．このような正解ラベルごとの誤答傾向については 3.2 節にまとめている．

2.2 非関連度スコアの導入

非関連度に着目した評価事例 LLM が関連度を過大評価する事例に対処するため，文書がクエリに"関連しない"理由に注目して評価を試みた．具体的には，図 8 に示したプロンプトを用いてスコアを出力させ，さらに LLM の出力の根拠を確認するため，スコアの理由を追加生成させた．図 1 と同様のクエリ–文書に対し出力させた結果を図 2 に示す．関連しない理由が適切に述べられているほか，"関連しない度数"は一貫して 2–3 点であり，関連度スコアよりも適切に評価できていることがわかる．

提案：非関連度スコアによるリランキング 上記の結果を受け，図 8 のようなプロンプトで出力させた"関連しない度数"に対して，式 2 で出力確率による重み付けを行ったスコアを**非関連度スコア**と定義

Dataset: news, **QID:** 882, **DocID:** db48dcdcc093c6fc5d5989d1359aa2c2
LLM non-relevance label and rationale (GPT-4o-mini): 3 out of 3. The passage primarily discusses the challenges faced by academic medical centers and does not mention potential medical breakthroughs or their origins from unexpected industries.

図 2: 関連しない理由の出力例.

する.

$$S_n(q,d) = \sum_{k=0}^{N_n} k \frac{\exp(z_k(q,d))}{\sum_{j=0}^{N_n} \exp(z_j(q,d))} \tag{2}$$

ここで $S_n(q,d)$ はクエリ q，文書 d に対する重み付けされた非関連度スコアであり，N_n は非関連度評価における最大整数ラベルである．非関連度スコアは関連度スコアに対し相反的な関係にあり，リランキングの際は昇順で並び替えて使用する．

3 実験

3.1 条件

ベースライン 評価には，TREC Deep Learning 2019/2020(DL19/DL20) および BEIR の一部データセット (TREC–COVID, NFCorpus, Touche–2020, DBPedia, SciFact, Signal-1M, TREC–NEWS, Robust04) を用いた．候補文書の生成には BM25 を用い，クエリごとに上位 100 件を取得し，リランキングの対象とした．ベースラインとするリランキング手法は関連度スコアを用いた pointwise 手法とし，図 7 のプロンプトおよび式 1 を用いた．これに対し，提案する手法は非関連度スコアを用いた pointwise 手法とし，図 8 のプロンプトおよび式 2 で算出し，昇順でリランキングした．

LLM 設定 LLM には GPT-4o-mini および Llama-3.1-8b-Instruct を使用した．推論条件として GPT-4o-mini は temperature を 1.0 とし，seed を固定して，各クエリ–文書対について 1 回の推論でそれぞれのスコアを取得した．Llama モデルでは，入力末尾における次トークンの logits を取得し，temperature を 1.0 とした softmax により確率分布を算出した．入力が長い場合には，全体のプロンプトがモデルの最大コンテキスト長 (128,000 トークン) に収まるよう文書の先頭を切り出した．

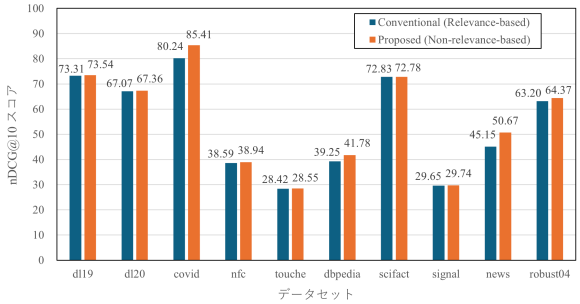

図 3: nDCG@10(gpt-4o-mini)

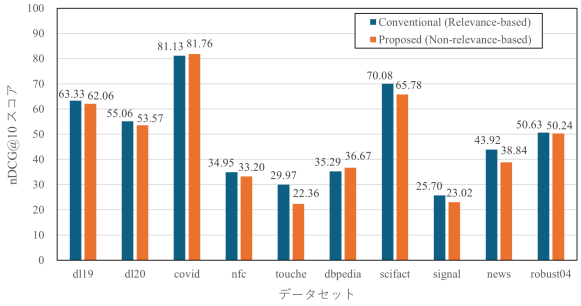

図 4: nDCG@10(Llama-3.1-8b-Instruct)

評価指標 評価指標として nDCG@10 を用い，pytrec eval により算出した [11]．統計的有意差は，提案手法とベースラインのクエリごとの nDCG@10 差を計算し，クエリ単位の bootstrap resampling を 10,000 回行って差分の 95%信頼区間を推定した．95%信頼区間が 0 を跨がない場合に有意差ありと判定した．

3.2 結果と分析

nDCG@10 スコア 図 3，4 は各データセットにおける nDCG@10 スコアを表す．GPT モデルを使用した場合，非関連度を用いた手法が従来手法と同等もしくは高いスコアを示していることがわかる．一方，Llama モデルでは多くのデータセットにおいてスコアが低下していることが見て取れる．

有意差 図 5，6 は各データセットにおける差分スコア (提案手法-従来手法) とその 95%信頼区間を図示したものである．黒点はクエリごとの差分スコアの平均値を示し，エラーバーはクエリ単位のブートストラップにより算出した 95%信頼区間を示す．GPT モデルでは TREC–COVID，DBPedia，TREC–NEWS のデータセットにおいて，スコア差の信頼区間が 0 を跨がず，提案手法が従来手法より

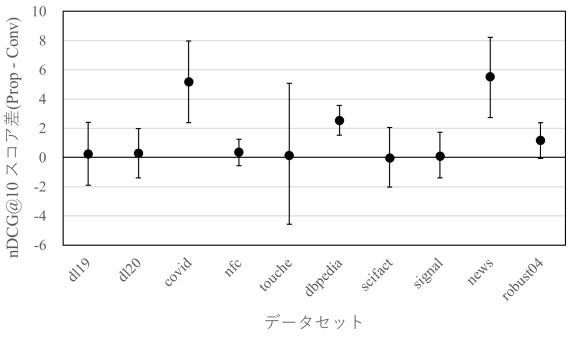

図 5: nDCG@10 スコアの有意差 (gpt-4o-mini)

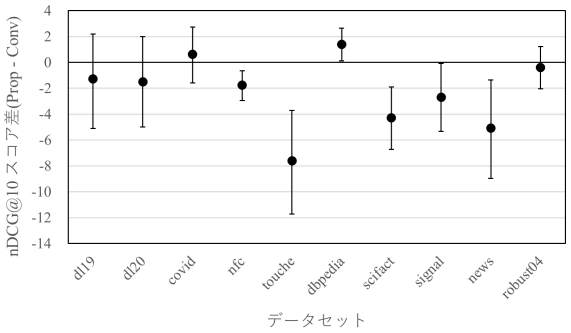

図 6: nDCG@10 スコアの有意差 (Llama-3.1-8b-Instruct)

統計的に有意に改善したことが確認できる．一方，Llama モデルでは有意に改善したのは DBPedia のみであり，Touche–2020，Signal-1M，TREC–NEWS，NFCorpus，SciFact の 5 つのデータセットで有意に悪化した．

分析　非関連度スコアによりスコアリングがどのように改善したのか分析するため，正解ラベルと関連度スコアおよび非関連度スコアを 0–1 に正規化し，文書ごとの正解ラベルとの絶対誤差をそれぞれ求めたのち，データセット全体における平均値を計算した．正規化において，各スコアは 1 に近いほど関連度が高くなるように正規化した．分析対象のデータセットは，GPT および Llama 両モデルにおいて有意な改善が見られ，かつクエリ数が 400 と十分大きな DBPedia を対象とした．表 1 に各モデルの正解ラベルごと平均絶対誤差をまとめた．

表 1 より，両モデルにおいて正解ラベル 2 の場合に精度が大きく向上したことが確認できる．これは，図 1 のように LLM が関連度を過大評価したケースに対し，非関連度スコアを用いることで適切な評価ができるようになったことを示している．一方，正解ラベル 0，1 の場合は非関連度スコアの精度が悪化している．ここで正解ラベル 0 に該当する非関連度スコアは，図 8 における最大スコア 3 点に対応している．このことから，非関連度スコアを用いる場合は LLM が非関連度を過大評価する傾向が推察される．

表 1: DBPedia における正解ラベルごと平均絶対誤差

	正解ラベル	関連度スコア	非関連度スコア
gpt-4o-mini	0	0.180	0.277
	1	0.364	0.380
	2	0.169	0.085
Llama-3.1-8b-Instruct	0	0.322	0.387
	1	0.188	0.357
	2	0.372	0.203

また，GPT モデルを使用した場合は正解ラベル 0，2 の精度が比較的高いのに対し，Llama モデルでは中間の正解ラベル 1 の精度が高い傾向がある．このことは，GPT モデルが最低点や最高点といった評価を下しやすい一方で中間点を出力しづらいこと，逆に Llama モデルでは中間的なスコアを出力しやすいことを示唆している．pointwise スコアリングに LLM を使用する際は，こういったモデルの特徴を考慮して使用する必要がある．

以上の特徴はデータセットによって異なる可能性があり，さらなる分析が必要である．特に Llama モデルを使用した場合は非関連度スコアによって有意に精度が悪化する場合が多く，使用には注意が必要である．

4　おわりに

本研究では，LLM を用いた pointwise リランキング手法の精度向上のため，非関連度スコアの導入を提案した．まず従来の関連度に注目したスコアリングの傾向を分析し，LLM が関連度を甘く評価する事例があることを報告した．そこで非関連度に注目したスコアリングを導入することで，より適切な評価が可能になり，特に GPT-4o-mini を使用した場合に複数のデータセットで精度が有意に向上することが確認された．一方，非関連度スコアを用いることで LLM が非関連度を過大評価する傾向も明らかになった．今後はさらに包括的な検証に加え，関連度スコアや非関連度スコアの傾向に合わせた活用方法の検討を進めていく．

参考文献

[1] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? Investigating large language models as re-ranking agents. **arXiv preprint arXiv:2304.09542**, 2023.

[2] Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. FIRST: Faster improved listwise reranking with single token decoding. **arXiv preprint arXiv:2406.15657**, 2024.

[3] Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 38–47, 2024.

[4] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. In **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 1504–1518, 2024.

[5] Fang Guo, Wenyu Li, Honglei Zhuang, Yun Luo, Yafu Li, Le Yan, Qi Zhu, and Yue Zhang. MCRanker: Generating diverse criteria on-the-fly to improve pointwise LLM rankers. In **Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining**, pp. 944–953, 2025.

[6] Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In **Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 654–664, 2021.

[7] Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In **Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (volume 2: short papers)**, pp. 358–370, 2024.

[8] Charles Godfrey, Ping Nie, Natalia Ostapuk, David Ken, Shang Gao, and Souheil Inati. Likert or not: LLM absolute relevance judgments on fine-grained ordinal scales. **arXiv preprint arXiv:2505.19334**, 2025.

[9] Nour Jedidi, Yung-Sung Chuang, James Glass, and Jimmy Lin. Don't" overthink" passage reranking: Is reasoning truly necessary? **arXiv preprint arXiv:2505.16886**, 2025.

[10] Xuan Lu, Haohang Huang, Rui Meng, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. Rethinking reasoning in document ranking: Why chain-of-thought falls short. **arXiv preprint arXiv:2510.08985**, 2025.

[11] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. **ACM Transactions on Information Systems (TOIS)**, Vol. 20, No. 4, pp. 422–446, 2002.

A　プロンプト

図 7 および図 8 は，関連度スコアおよび非関連度スコアを出力するために実際に使用したプロンプトである．

```
You are an expert evaluator for information
    retrieval (IR) systems.
Your task is to evaluate **how relevant a
    passage is to a given query**,
based on whether the passage contains
    information that could directly or
    indirectly answer the query.

Please output **only one integer (0-3)**
    according to the following scale:

3 = HIGHLY_RELEVANT
- Fully satisfies the main information need.
- Contains detailed, specific, and directly
    useful information.
- Provides substantial value beyond a simple
    mention.

2 = RELEVANT
- Addresses the information need meaningfully.
- Provides some useful information, but may
    lack depth or completeness.
- More than a superficial mention; still
    clearly on-topic.

1 = PARTIALLY_RELEVANT
- The document touches the topic but only
    superficially.
- Contains limited or tangentially useful
    information.
- Provides minor value to the user.

0 = NOT_RELEVANT
- Does not address the information need.
- Only contains coincidental keyword matches OR
     is on a different topic.

Output format rule:
- Output **only the number (0-3)**. No words,
    punctuation, or explanations.

query:
passage:
```

図 7: 関連度スコア出力用プロンプト．

```
You are an expert evaluator for information
    retrieval (IR) systems.
Your task is to evaluate **how unrelated a
    passage is to a given query**.
Focus only on the degree to which the passage
    fails to provide information that could
    answer the query directly or indirectly.

Please output **only one integer (0-3)**
    according to the following scale:

3 = COMPLETELY_UNRELATED
- No information that helps answer the query.
- Different topic, context, or domain.
- No meaningful conceptual connection.

2 = MOSTLY_UNRELATED
- Only minor or coincidental overlap (e.g.,
    shared keywords).
- Does not contribute useful information toward
     answering the query.

1 = PARTIALLY_UNRELATED
- Some connection exists, but insufficient for
    answering the query.
- Relevance is indirect, partial, or minimal.

0 = NOT_UNRELATED
- Contains clear and meaningful information
    that supports answering the query.
- Cannot be considered unrelated.

Output format rule:
- Output **only the number (0-3)**. No words,
    punctuation, or explanations.

query:
passage:
```

図 8: 非関連度スコア出力用プロンプト．