

논문 2026-3-20 <http://dx.doi.org/10.29056/jsf.2026.09.20>

금융공시 RAG의 연도 혼동에 따른 수치 질의 오검색: 원인 분석과 완화 방안

손정헌*, 장효경**†

Year Confusion in Financial RAG: Understanding and Mitigating Wrong-Year Retrieval for Numerical Queries

Jeongheon Shon*, Hyokyung Chang**†

요 약

검색 증강 생성(Retrieval-Augmented Generation, RAG)에서 검색 오류는 생성 모델이 잘못된 근거를 참고하게 한다. 금융공시는 동일 기업의 유사 항목과 비교수치가 여러 연도에 반복되어, 질문이 요구한 연도와 다른 공시가 목표 공시보다 높은 순위로 검색될 수 있다. 본 연구는 이를 ‘연도 혼동’으로 정의하고, 후보 문서 수, 기업 관계, 연도 거리, 내용 중복에 따른 목표 공시의 순위 변화를 분석하였다. 후보 수를 같게 맞춘 비교에서도 BM25의 목표 공시 선택률은 동일 기업·인접 연도 조건에서 타 기업·같은 연도 조건보다 낮았다. 연도 혼동을 완화하기 위해 기존 검색 결과에 질문과 공시의 보고기간 관계를 반영하는 Temporal 재순위화 프레임워크를 제안하였다. 주요 평가 집합 809개에서 Temporal은 HR@1 0.7157, MRR@10 0.8463을 기록했으며, HR@1은 BM25보다 12.11%p 향상되었다. 이 결과는 보고기간 정보가 내용 관련성을 보완하여 목표 공시 선택 성능을 높였음을 보여준다.

Abstract

Retrieval errors in retrieval-augmented generation (RAG) cause generation models to rely on incorrect evidence. Financial filings repeat similar items and comparative figures for the same company across years, allowing filings from other years to outrank the target filing. We defined this as “year confusion” and analyzed target filing ranks by candidate count, company relationship, year distance, and content overlap. With equal candidate counts, BM25 selected the target filing less often under the same-company, adjacent-year condition than under the different-company, same-year condition. We proposed a Temporal reranking framework that incorporates reporting-period relationships between questions and filings into existing retrieval results. On 809 primary evaluation cases, Temporal achieved HR@1 of 0.7157 and MRR@10 of 0.8463, exceeding BM25 by 12.11 percentage points in HR@1. These results show that reporting-period information complements content relevance and improves target filing selection.

한글키워드 : 검색 증강 생성, 금융 공시, 연도 혼동, 수치 질의, 문서 재순위화

keywords : retrieval-augmented generation, financial filings, year confusion, numerical queries, document reranking

* 한남대학교 컴퓨터공학과

** 한남대학교 컴퓨터공학과

† 교신저자: 장효경(email: chantellejang@hnu.kr)

접수일자: 2026.09.02. 심사완료: 2026.09.11.

게재확정: 2026.09.20.

1. 서론

외부 문서를 검색해 생성 모델의 답변 근거로 넘기는 검색 증강 생성(Retrieval-Augmented Generation, RAG)에서는 검색기가 무엇을 고르느냐가 답변의 출발점을 결정한다[1]. 질문과 관련 있어 보이지만 보고기간이 다른 공시가 선택되면, 생성 모델은 그 문서의 수치를 맞는 근거로 받아들일 수 있다. 금융 질의에서는 이 위험이 특히 크다. 동일 항목명이나 유사한 숫자가 반복되더라도 기업·보고기간·단위가 다르면 전혀 다른 사실을 가리키기 때문이다. 본 연구는 답변 생성에 앞선 문서 선택 단계에 초점을 맞춰, 질문이 요구한 시점과 다른 보고기간의 공시가 목표 공시보다 높은 순위에 놓이는 현상을 분석한다(그림 1).



그림 1. 금융공시 연도 혼동에 따른 목표 공시의 순위 역전

Fig. 1. Target-filing rank reversal caused by year confusion in financial retrieval

기업의 연차공시는 매년 처음부터 다시 쓰는 문서가 아니다. 회계항목과 표의 틀이 이어지고, 앞선 연도의 비교수치가 다음 연도 공시에 반복해서 들어간다[5][6]. 가령 2020년 공시에 2019년 비교수치가 실려 있다면, 2019년 수치를 묻는 질문은 2019년 공시뿐 아니라 2020년 공시와도 내용상 겹친다. 이때 후속 연도의 공시가 내용 유사성 때문에 목표

공시(target)보다 높은 순위에 오를 수 있다. 이처럼 질문이 요구한 연도와 다른 공시가 목표 공시의 순위를 앞서는 현상을 ‘연도 혼동(year confusion)’이라 정의한다.

후보 공시의 수가 같더라도 기업 관계, 보고연도 거리와 내용 중복에 따라 목표 공시의 순위는 달라질 수 있다. 본 연구는 후보 수와 이들 조건에 따른 순위 변화를 비교하여 연도 혼동이 두드러지는 상황을 분석한다. 이어 기존 검색 관련성에 질문과 공시의 보고기간 관계를 결합한 Temporal 재순위화 방법을 제안하고, 목표 공시 선택의 개선 여부를 검증한다.

2. 연구 배경 및 관련 연구

이 절에서는 금융 수치 QA와 장문 공시 검색의 특성을 살펴보고, 검색 기술과 장문 문서 처리, 조건 불일치 후보, 시간·버전 정합성을 다룬 관련 연구를 차례로 검토한다.

2.1 금융 수치 QA와 장문 공시 검색

Chen et al.의 FinQA는 표와 문장이 함께 제시된 금융 보고서에서 수치 질문을 풀도록 만든 데이터셋이다[2]. 질문과 정답만 주는 데 그치지 않고, 계산에 사용되는 문장·표와 연산 순서도 함께 제공한다. 이를 통해 모델의 정답 여부뿐 아니라 근거 선택과 계산 과정도 확인할 수 있다.

FinQA에서 질문을 푸는 데 필요한 문맥은 비교적 짧다. 따라서 전체 공시를 훑어 필요한 근거의 위치를 찾는 검색 문제를 직접 평가하기에는 범위가 제한된다. Reddy 등(2024)의 DocFinQA는 이러한 한계를 보완하기 위해 FinQA의 7,437개 질문을 각 질문의 출처인 전체 SEC 공시와 연결하여, 짧은 문맥을 넘어 장문 금융문서 안에서 필요한 근거를 찾아야 하는 환경을 제공한대[3]. 다만

공개 데이터는 각 질문과 하나의 출처 공시를 연결하는 구조로 구성되어 있다.

2.2 검색 기술과 문서 처리

BM25는 질문과 문서에 함께 등장하는 단어나 숫자를 바탕으로 관련성을 계산한다[9]. 금융문서 처럼 지표명과 수치 표현이 중요한 환경에서는 직접적인 어휘 일치에 유용한 검색 신호가 되지만, 같은 의미가 다른 표현으로 나타나면 단어의 중복만으로 관련성을 판단하기 어렵다. Dense retrieval은 질문과 문서를 의미 벡터로 변환한 뒤 벡터 간 유사성을 비교하여 이러한 한계를 보완한다[10][12].

장문 문서는 임베딩 모델의 입력 길이를 넘어설 수 있으므로, 문서를 여러 청크나 window로 나누어 처리하는 방식이 널리 사용된다. 이 경우 각 구간에서 계산된 관련성을 다시 문서 수준으로 모으는 과정이 필요하다. 서로 다른 검색 결과를 결합할 때에는 Reciprocal Rank Fusion(RRF)을 사용할 수 있다[11]. RRF는 검색기마다 점수의 범위가 다르더라도 원점수를 직접 맞추지 않고 각 검색 결과에서의 순위를 이용해 결과를 결합한다.

2.3 금융 검색의 조건과 방해 후보

검색 과정에서는 질문과 전혀 무관한 문서뿐 아니라, 질문과 많은 표현을 공유하면서도 필요한 조건을 충족하지 않는 문서가 정답 선택을 방해할 수 있다. 이러한 경쟁 문서를 방해 후보(distractor)라 하며, 그중 실제 정답 근거가 아니면서 목표 공시와 내용상 매우 유사하여 구별하기 어려운 후보를 hard negative로 구분한다[6]. 따라서 hard negative는 단순히 기업이나 기간이 다르다는 사실만으로 정해지는 것이 아니라, 정답 여부와 내용 유사성을 함께 고려하는 개념이다.

Mansouri 등(2026)의 FinRank는 미국 증권거래위원회 제출 공시에서 정답 근거와 내용상 유사한 hard negative를 함께 제시하고, 검색기가 이들을 얼마나 잘 구별하는지를 평가한다[6]. 같은 기업의 다른 연도 공시와 유사 기업의 공시를 포함함으로써, 단순한 어휘 일치만으로는 정답을 가리기 어려운 검색 환경을 구성하였다.

금융 검색의 조건을 보다 직접적으로 반영하려는 접근도 이어졌다. Jiang 등(2026)의 Fin-RATE는 질문이 가리키는 기업과 연도를 먼저 식별한 뒤 그 범위 안에서 관련 내용을 찾는 단계적 검색 방식을 제안하였다[5]. Zhou 등(2026)의 FinCARDS는 기업, 금융항목, 회계기간, 실제 수치 등의 정보를 카드 형태로 구조화하고, 질문 카드와 후보 카드의 일치 정도를 이용해 검색 결과를 다시 정렬하였다[4]. 이처럼 어휘나 의미의 관련성뿐 아니라 기업, 기간, 항목과 같은 조건을 검색 과정에 구조적으로 반영하는 방법이 연구되어 왔다.

2.4 시간 · 버전 정합성

문서가 시간에 따라 갱신되거나 여러 버전으로 존재하는 환경에서는 검색 결과의 내용뿐 아니라 어느 시점의 문서인지를 함께 고려해야 한다. Huwiler 등(2025)의 VersionRAG는 변화하는 문서의 여러 버전 가운데 질문에 적합한 버전을 검색하는 문제를 다루었다[7].

법률 문서에서도 시점이 다른 버전을 잘못 근거로 선택하는 문제가 관찰된다. Cymbler 등(2026)은 프랑스 세법의 버전별 코퍼스를 이용해, 내용상 관련된 문서라도 질문에 필요한 시점과 맞지 않으면 잘못된 근거가 될 수 있는 temporal misgrounding을 분석하였다[8]. 이러한 연구는 내용의 관련성과 시간적 적합성이 서로 다른 검색 조건이 될 수 있음을 보여준다.

3. 데이터와 제안 프레임워크

3.1 데이터 구성

실험에는 DocFinQA의 질문 - 공시 연결 자료 [3]를 사용하였다. 전체 7,437개 질문 가운데 기업을 구분할 임시 식별자를 결정할 수 없는 3개 Context와 연결된 질문 6개를 제외하여 7,431개를 사용하였다. 기업 식별값은 FinQA 파일명에서 추출한 단서를 바탕으로 만든 provisional ticker로 정하였다. 이는 SEC가 제공하는 공식 ticker가 아니라 기업 단위 데이터 분할을 위한 내부 식별값이다.

같은 기업의 질문이 학습과 평가에 동시에 포함되는 것을 막기 위해 provisional ticker를 기준으로 기업 단위 split을 구성하였다. 총 124개의 ticker 그룹을 train, validation, test로 나누고, 하나의 기업이 서로 다른 split에 중복되지 않도록 하였다. 이후 모든 학습과 평가에는 이 분할을 그대로 유지하였다.

주요 평가 집합(Primary)은 test의 1,193개 질문에서 단계적으로 선정하였다. 이 가운데 질문에 명시적인 대상 연도가 있는 982개를 먼저 추렸고, 다시 같은 기업의 인접 연도 방해 후보를 구성할 수 있으면서 DocFinQA가 연결한 목표 공시가 하나로 정해지는 질문을 남겼다. 최종적으로 809개가 이 조건을 만족했으며, 이 가운데 후보가 2개인 질문은 140개, 3개인 질문은 669개였다.

선택 순서를 고정하여 후보군을 다시 구성해도 같은 문서가 선택되도록 구성하였다. 각 후보군에는 목표 공시 1개와 관계 조건에 따라 고른 방해 후보를 포함하며, 방해 후보는 목표 공시와 같은 split에서만 선택하였다. 동일한 Context가 중복되지 않도록 해시값을 비교하였다. 동일 기업 후보에서는 공시 헤더의 보고연도 차이가 1년이면 인접 연도, 2년 이상이면 먼 연도로 구분하

고, 같은 조건의 후보가 여러 개일 때는 Context 해시 순서로 결정하였다. 타 기업 후보의 선택 순서는 고정된 seed, 질문 ID, 관계 조건과 Context 해시를 결합하여 정하였다.

관계 조건에 따라 후보군을 구성할 수 있는 질문 수와 후보 문서 수는 서로 달랐다. 조건별 가용 질문 수와 후보군 크기는 표 1에 정리하였다. 개별 관계 조건은 목표 공시와 해당 관계의 방해 후보로 구성하였고, Mixed는 여러 관계의 방해 후보를 함께 포함하였다.

표 1. 관계 조건별 가용 질문 수와 후보군 크기

Table 1. Number of available queries and candidate-set size by relation condition

후보 조건	가용 질문 수	후보군 크기(K)
Single	1,193	1
동일 기업·인접 연도	962	2~3
동일 기업·먼 연도	1,142	2~3
타 기업·같은 연도	1,132	2~3
타 기업·다른 연도	1,177	2~3
Mixed	1,177	5~9

Single은 목표 공시만 포함하므로 HR@1이 구조적으로 1이 된다. 따라서 이를 성능 기준선으로 해석하지 않고, 방해 후보가 추가되었을 때 목표 공시의 순위가 어떻게 달라지는지를 확인하기 위한 비교 조건으로 사용하였다.

여기서 평가하는 것은 전체 SEC 코퍼스에서 관련 공시를 처음부터 찾는 Stage-1 검색이 아니다. 목표 공시가 포함된 후보군을 사전에 구성한 뒤 그 안에서 문서 순위를 다시 정하는 통제된 재순위화 문제를 평가한다.

3.2 제안의 출발점: 내용 관련성과 보고기간 정합성

다른 연도의 공시가 높은 검색 점수를 받았다고 해서 내용 검색 자체가 실패했다고 볼 수는 없다.

후속 공시가 질문의 금융지표와 과거 수치를 실제로 포함한다면, 그 문서는 내용 측면에서도 충분히 관련될 수 있기 때문이다. 문제는 이렇게 관련성이 높은 여러 공시가 함께 놓였을 때 질문과 연결된 목표 공시를 가려내지 못하는 데 있다.

내용 관련성과 시간적 적합성을 구분한다는 점은 시간·버전 검색 연구의 문제의식[7][8]과 맞닿아 있다. Temporal은 기존 검색 관련성에 질문과 공시의 보고기간 관계를 별도의 순위 신호로 결합한다.

연도가 다른 후보를 제거하는 hard filter는 사용하지 않았다. 후속 공시의 과거 비교수치나 여러 시점을 비교하는 질문의 근거까지 제외할 수 있기 때문이다. 대신 후보를 유지한 상태에서 보고기간 적합성을 최종 순위에 반영하는 soft reranking을 사용하였다.

3.3 제안 프레임워크의 구조

조건별 후보군 구성과 보고기간 기반 문서 재순위화의 전체 흐름은 그림 2와 같다.

각 공시는 먼저 2,750자 길이의 청크로 나누고, 인접 청크 사이에는 550자의 중첩을 두었다. BM25는 이 청크를 단위로 어휘와 수치의 일치를 계산한다. Hierarchical MaxP는 각 청크를 다시 content 254 token, overlap 64 token의 subwindow로 분할하며, 특수 token을 포함한 입력 길이는 최대 256 token으로 제한하였다. 질문과 각 subwindow의 의미 유사도를 계산한 뒤 가장 높은 값을 해당 청크의 MaxP 점수로 사용하였다.

비교 방식인 Direct512는 각 2,750자 청크를 최대 512 token까지 입력하여 질문과의 의미 유사도를 계산하고, 문서에 속한 청크별 점수 가운데 최댓값을 해당 문서의 점수로 사용하였다. 입력 길이에는 특수 token을 포함하였다.

BM25와 Hierarchical MaxP가 만든 청크 순위는 Equal RRF($k=60$)로 결합하였다. 이후 각 공시

에 속한 청크 가운데 가장 높은 RRF 점수를 선택하여 문서 수준의 기본 검색 점수를 만들었다. 시간 특징을 계산할 때에는 각 문서에서 두 검색기의 상위 10개 청크를 합친 뒤, 동일한 청크 ID는 한 번만 남겼다. 이렇게 검색 단계에서 얻은 정보를 문서 단위로 정리한 뒤 보고기간 관련 특징과 결합하였다.

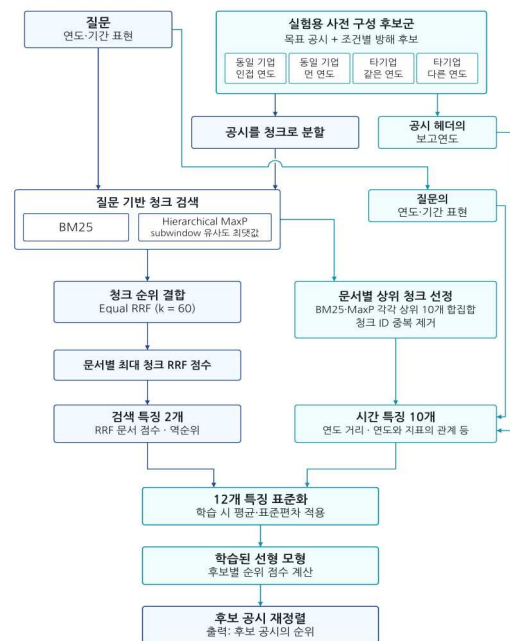


그림 2. 조건별 후보군과 보고기간 기반 문서 재순위화의 구조

Fig. 2. Condition-specific candidate pools and reporting-period-aware document reranking pipeline

Temporal의 재순위화 입력은 기존 검색 관련성을 나타내는 검색 특징 2개와 질문 - 공시 간 보고기간 관계를 나타내는 시간 특징 10개로 구성하였다. 검색 특징은 Equal RRF 점수와 역순위이며, 시간 특징은 질문의 연도 표현, 후보 공시의 보고연도, 질문 연도와 공시·본문 정보의 관계를 서로

다른 수준에서 나타낸다. 각 시간 특징의 정의는 표 2에 정리하였다.

표 2. Temporal 재순위화에 사용하는 시간 특징
Table 2. Temporal features used for document reranking

시간 특징	정의
질문 내 연도 수	질문에 포함된 연도 표현의 수
후보 공시 헤더 연도	후보 공시의 보고연도
최소 절대 연도 거리	질문 연도와 헤더 연도의 최소 절대 거리
부호 연도 거리	질문 연도와 헤더 연도의 방향을 포함한 거리
질문 연도 출현 비율	질문 연도가 후보 문서에 나타나는 비율
연도 - 지표 동일 표	연도와 금융지표가 같은 표에 나타나는지 여부
연도 - 지표 인접 행	연도와 금융지표가 인접 행에 나타나는지 여부
연도 - 지표 동일 window	연도와 금융지표가 같은 window에 나타나는지 여부
인접연도 지배도	인접 연도의 단서가 상대적으로 우세한 정도
상대기간 표현 양립성	질문과 후보 공시의 상대기간 표현이 양립하는 정도

공시의 보고연도가 질문 연도와 가깝다는 것과 본문에서 해당 연도가 금융지표와 함께 제시된다는 것은 서로 다른 단서다. 시간 특징은 헤더 수준의 연도 관계뿐 아니라 본문에서의 제시 맥락도 나타내도록 구성하였다.

이 특징들을 표준화한 뒤 L2 정규화를 적용한 logistic regression으로 결합하였다. 후보 공시의 순위는 모형의 선형 결합값으로 정하므로, 출력은 정답 확률이 아닌 선형 순위 점수가 된다. 모형은 train에서 학습하였으며, validation에서 rrf_temporal과 C=1.0으로 설정을 고정한 뒤 test에 적용하였다. test 결과를 확인한 뒤에는 설정을 다시 조정하지는 않았다.

정보 누수를 막기 위해 정답 여부나 목표 공시 역할을 직접 드러내는 답(answer), 계산 과정(program), gold 근거 인덱스(gold_inds), 후보의 target 역할, 검토 결과(review verdict) 및 목표 공시 ID 같은 정보는 추론 입력에서 제외하였다.

3.4 기존 방식과의 비교 설계

보고기간 정보의 추가가 목표 공시 선택을 개선하는지 확인하기 위해, Temporal을 단일 검색 방식 및 일반적인 hybrid 결합과 비교하였다. 기준선은 어휘 기반 검색, 장문 dense 검색, 입력 길이를 제한한 dense 검색과 두 검색기의 순위 결합으로 구성하였다. 각 방식이 사용하는 정보와 비교에서의 역할은 표 3과 같다.

표 3. 비교 방식이 사용하는 정보와 역할
Table 3. Information and role of the compared retrieval methods

방식	주요 정보	역할·한계
BM25	단어·숫자 중복	강한 lexical 기준선
Hier. MaxP	subwindow 의 미 유사도	장문 dense 기준선
Direct512	청크당 최대 512 token	입력 길이 제한을 둔 dense 기준선
Equal RRF	BM25·MaxP 청크 순위	일반 hybrid 기준선
Temporal	검색 특징+시간 특징	제안 재순위화

기업 관계에 따른 차이를 비교할 때에는 같은 질문에 대해 K=3 후보군을 구성할 수 있는 경우만 대응시켰다. 동일 기업·인접 연도와 타 기업·같은 연도의 비교에는 725개 질문을, 동일 기업·먼 연도와 타 기업·다른 연도의 비교에는 1,114개 질문을 사용하였다. 이 설계는 질문과 후보 수를 맞추지만, 후보의 업종, 금융지표, 문서 길이와

어휘·의미 유사도까지 통제하지는 않는다. 따라서 이 비교는 기업 관계만의 독립적인 효과가 아니라, 해당 조건으로 구성된 후보군 사이의 차이를 보여준다.

후보 수의 영향은 같은 질문 안에서 $K=3$ 과 $K=2$ 를 비교하였다. 각 관계 조건에서 원래 $K=3$ 이었던 질문만 사용하고, $K=2$ 에서는 목표 공시와 선택순위 1의 방해 후보만 남겼다. 두 조건 모두 BM25와 MaxP의 청크 순위를 RRF로 결합한 뒤 문서별 최고 점수를 사용하는 동일한 계산 절차를 적용하였으며, 시간 특징과 학습 모형도 그대로 고정하였다. 이 비교에 포함된 원래 $K=3$ 질문은 동일 기업·인접 연도 790개, 동일 기업·면연도 1,114개, 타 기업·같은 연도 1,108개, 타 기업·다른 연도 1,177개이다.

주요 평가 지표는 Primary의 strict HR@1로 정하였다. 여기서 strict target은 DocFinQA가 연결한 목표 공시를 뜻하며, strict HR@1은 이 문서가 후보 가운데 1위에 놓인 질문의 비율이다. MRR@10은 목표 공시가 전체적으로 어느 순위에 위치하는지를 함께 보기 위한 보조 지표로 사용하였다. wrong-period rank-1은 목표 공시와 다른 보고연도의 공시가 1위에 놓인 비율이다. Primary에서는 모든 방해 후보가 동일 기업의 인접 연도 공시이므로 wrong-period rank-1은 $1 - \text{strictHR@1}$ 과 같아진다. 이 값은 문서 선택 단계에서 보고기간이 어긋난 비율이며, 최종 수치 답변이나 공시 내부 근거의 오류율을 의미하지 않는다.

Primary의 주요 성능 차이에 대한 불확실성은 평가 split의 ticker 18개를 cluster로 두고 bootstrap으로 계산하였다. 각 반복에서 ticker를 복원추출하고, 선택된 ticker에 속한 Primary 질문을 모아 성능 차이를 다시 계산하였다. 이를 10,000회 반복하여 percentile bootstrap 95% 신뢰구간을 구하였다(seed=20260826). 평가 split의

18개 ticker 가운데 실제로 Primary 질문을 제공한 그룹은 11개였고, 나머지 7개에는 Primary 질문이 없었다.

Primary의 사전 지정 주요 비교는 Temporal-BM25와 Temporal-Equal RRF 두 쌍으로 정하였다. 각 비교에는 질문별 strict HR@1 성공 여부에 대한 양측 exact McNemar 검정을 적용하고, 두 p값은 Holm 방법으로 보정하였다. Temporal의 성능 개선은 세 조건을 모두 만족할 때 확인된 것으로 판단하였다. 첫째, Temporal의 성능 점추정치가 비교 방식보다 높아야 한다. 둘째, 성능 차이에 대한 ticker-cluster bootstrap 95% 신뢰구간이 0을 포함하지 않아야 한다. 셋째, Holm 보정 후 p값이 0.05보다 작아야 한다.

4. 실험 결과 및 해석

4.1 주요 재순위화 성능과 완화

Primary 809개에서 비교한 다섯 방식 가운데 Temporal은 가장 높은 strict HR@1 0.7157과 MRR@10 0.8463을 기록하였다. 시간 특징을 사용하지 않은 기준선 중에서는 BM25의 HR@1이 0.5946으로 가장 높았으며, Equal RRF는 0.4932였다. 각 방식의 HR@1, MRR@10과 wrong-period rank-1은 표 4에 제시하였다.

표 4. Primary strict 문서 재순위화 결과
Table 4. Primary strict document-reranking results

시스템	HR@1	MRR@10	wrong-period rank-1
BM25	0.5946	0.7789	0.4054
Hier. MaxP	0.4190	0.6722	0.5810
Direct512	0.3869	0.6545	0.6131
Equal RRF	0.4932	0.7241	0.5068
Temporal	0.7157	0.8463	0.2843

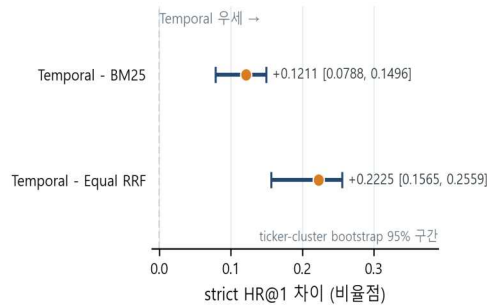


그림 3. Temporal 재순위화의 strict HR@1 개선과 95% bootstrap 구간

Fig. 3. Strict HR@1 gains from temporal reranking with 95% bootstrap intervals

Temporal의 HR@1은 BM25보다 12.11%p, Equal RRF보다 22.25%p 높았다(그림 3). 각 차이의 ticker-cluster bootstrap 95% 신뢰구간은 각각 [0.0788, 0.1496]과 [0.1565, 0.2559]로 모두 0을 포함하지 않았으며, exact McNemar 검정의 Holm 보정 p값도 각각 $8.409e-10$ 과 $3.766e-27$ 이었다. 다시 말해, 두 비교 모두 3.4에서 정한 세 가지 성능 개선 기준을 충족하였다.

4.2 후보 수와 기업 관계

두 번째 방해 후보를 추가하자 네 관계 조건 모두에서 HR@1이 하락하였다. 그러나 하락 폭은 후보의 관계에 따라 달랐다. BM25는 동일 기업·인접 연도에서 0.1506이 낮아진 반면, 타 기업·다른 연도에서는 0.0459가 낮아졌다. Temporal에서도 같은 방향이 나타났지만, 주 분석의 네 조건 모두에서 하락 폭은 BM25보다 작았다. 조건별 HR@1 변화량은 표 5에 정리하였다. 후보 수를 $K=3$ 으로 맞춘 비교에서도 관계에 따른 차이가 나타났다. BM25의 HR@1은 동일 기업·인접 연도 조건에서 타 기업·같은 연도보다 0.2386 낮았고, 동일 기업·먼 연도 역시 타 기업·다른 연도보다 0.1616 낮았다. 이 차이는 3.4에서 설명한 통제 범위 안에서 구성된 후보군 사이의 차이이며, 기업 동일성만의 독립

적인 효과를 의미하지 않는다.

표 5. 두 번째 방해 후보 추가에 따른 HR@1 변화($K=3-K=2$)

Table 5. HR@1 change after adding a second distractor ($K=3-K=2$)

후보 유형	n	BM25	Temporal
동일 기업·인접 연도	790	-0.1506	-0.1228
동일 기업·먼 연도	1,114	-0.0952	-0.0871
타 기업·같은 연도	1,108	-0.0794	-0.0758
타 기업·다른 연도	1,177	-0.0459	-0.0391

유지하는 방해 후보를 달리한 보조 분석에서는 타 기업·같은 연도 조건에서 반대 방향의 결과도 나타났다.

4.3 연도 혼동을 설명하는 관측 신호

동일 기업 후보에서는 보고연도 차이가 1년일 때 목표 공시와 방해 공시의 순위 역전이 가장 자주 나타났다. 4,008개의 target - distractor 쌍을 분석한 결과, 이 구간의 역전율은 BM25 0.3196, Temporal 0.2420이었다. 다만 연도 거리가 멀어질수록 역전율이 일관되게 낮아지는 단조 관계는 확인되지 않았다.

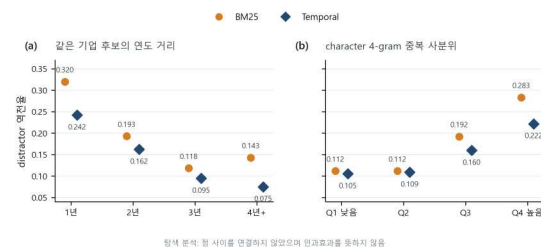


그림 4. 연도 거리와 내용 중복에 따른 방해 후보 역전율

Fig. 4. Exploratory distractor-outranking rates by year distance and content overlap

이러한 연도 거리의 차이가 실제 내용 중복과도 연결되는지 확인하기 위해, 질문 ID와 두 문서의 해시를 기준으로 중복을 제거한 8,602개의 target

-distractor 쌍을 분석하였다. 동일 기업 후보는 타 기업 후보보다 모든 중복 지표에서 높은 값을 보였다(표 6). 동일 기업 안에서도 character 4-gram 중복은 인접 연도가 0.7604로 먼 연도의 0.6804보다 높았다.

표 6. 기업 관계에 따른 내용 중복
Table 6. Content overlap by company relationship

중복 지표	동일 기업	타 기업
Character 4-gram	0.7153	0.5144
Token-set	0.6622	0.3514
숫자 token	0.2973	0.1158
Row-label	0.5920	0.1904

내용 중복의 크기와 순위 역전 사이에도 같은 방향의 관계가 나타났다. 중복도 사분위가 높아질수록 BM25와 Temporal 모두에서 방해 후보의 역전율이 증가하였고(그림 4), 중복도와 target-minus-distractor margin의 Pearson 상관계수도 각각 -0.3924 와 -0.3608 로 음의 값을 보였다. 즉 내용 중복이 클수록 목표 공시가 방해 공시보다 갖는 점수 우위가 작아지는 경향이 관찰되었다.

관찰된 결과는 동일 기업의 인접 연도 공시가 더 많은 내용을 공유하고 있고, 내용 중복이 큰 후보에서 목표 공시의 순위 역전이 더 자주 나타났다는 관계를 보여준다. 단, 동일 기업 여부와 보고연도 거리는 내용 중복과 함께 변하므로, 이 분석만으로 각 요인의 독립적인 인과효과를 분리할 수는 없다.

4.4 잔여 오류 분석

Temporal을 적용한 뒤에도 Primary 809개 가운데 230개에서는 strict target이 1위에 오르지 못했다. BM25에서 실패했다가 Temporal에서 성공한 질문은 177개였고, 반대로 BM25의 성공이 Temporal의 실패로 바뀐 질문은 79개였다. 결국 전

체 성능의 개선이 모든 질문의 개선을 뜻하지는 않았다.

타 기업·같은 연도 후보를 구성할 수 있었던 전체 1,132개에서 Temporal의 HR@1은 0.7385, BM25는 0.7580으로, Temporal이 원수치 기준 0.0194 (1.94%p) 낮았다. 따라서 Temporal의 상대적 우위가 모든 후보 관계에서 유지된 것은 아니다.

strict target 이외의 공시도 유효한 근거가 될 수 있어, 정답 인정 범위를 넓혔을 때의 결과를 추가로 비교하였다. 809개 질문과 검색 순위는 그대로 유지한 채 model-assisted blind review에서 alternative positive를 보조 정답으로 추가하자, HR@1은 Temporal 0.7392, BM25 0.6255로 계산되었다. 이 보조 민감도 분석에서도 Temporal의 HR@1은 BM25보다 높았다.

4.5 연구의 한계 및 향후 연구

본 연구에서는 DocFinQA가 각 질문과 연결한 Context를 strict target으로 삼아 문서 단위 재순위화 성능을 평가하였다[3]. 이는 가능한 모든 정답 근거를 전문가가 다시 판정한 gold에 대한 평가와는 구별된다. 4.4의 model-assisted review 역시 대체 근거의 존재 가능성을 살피는 민감도 분석으로, 독립적인 human gold를 대신하지 않는다. 향후 평가 정답의 범위를 보완하려면 두 명 이상의 금융문서 검토자가 근거를 독립적으로 판정하고 adjudication을 거쳐야 한다.

평가 결과의 일반화에는 질문의 범위와 경쟁 후보의 난도를 함께 고려해야 한다. Primary는 질문에 대상 연도가 명시되어 있고, 동일 기업의 인접 연도 공시를 방해 후보로 구성할 수 있는 경우로 한정하였다. 또한 포함된 질문은 모두 표를 근거로 답하거나 표와 본문을 함께 활용하는 유형에 해당하였다. 또한 타 기업 방해 후보는 업종, 금융지표, 문서 길이와 어휘·의미 유사도를 동일하게 맞춘 hard negative가 아니다. 향후에는 새로운 ticker를

포함한 독립 평가셋으로 대상을 넓히면서, 업종·지표·문서 길이와 의미 유사도를 맞춘 경쟁 후보에서도 일반화 성능을 검토해야 한다.

이번 평가는 목표 공시가 포함된 후보군 안에서의 재순위화에 한정되어, 전체 코퍼스 검색에서 발생할 수 있는 오류 전파를 직접 평가하지는 못하였다. Stage-1에서 목표 공시가 후보군에 포함되지 않으면 이후 재순위화만으로 이를 복구할 수 없으므로, 동일한 질문 집합과 strict target을 기준으로 평가할 때, 최종 HR@1은 1단계 검색 재현율(Recall@M)을 넘을 수 없다. 또한 관련 청크가 검색되지 않거나 헤더의 보고연도가 잘못 추출되면 시간 특징에도 오류가 전달되어 최종 순위에 영향을 줄 수 있다.

후속 실험에서는 문서 검색부터 근거 탐색과 답변 생성까지 평가 범위를 넓혀야 한다. Stage-1 Recall@M으로 목표 공시의 후보 포함 여부를 평가하고, 후보 내 HR@1로 재순위화 이후의 목표 공시 선택을 평가한다. 공시 내부의 근거 포함률과 최종 수치 답변 정확도도 각각 측정하여, 근거 탐색과 답변 생성 단계의 성능을 구분한다.

선택된 공시 내부의 근거 탐색에는 entity·metric·period·value를 구조화하는 FinCARDS와 같은 접근[4]을 결합할 수 있다. 향후에는 이러한 근거 탐색과 단계별 평가를 함께 수행하여, 문서 단위 재순위화의 개선이 실제 근거 선택과 최종 답변 성능으로도 이어지는지 검증할 필요가 있다.

5. 결론

본 연구는 여러 연도의 금융공시가 함께 경쟁하는 환경에서 연도 혼동을 분석하고, 기존 검색 관련성에 보고기간 정보를 결합한 Temporal 재순위화 방법을 제안하였다. 동일 기업의 인접 연도 공시는 높은 내용 중복과 순위 혼동을 보였으

며, Primary 809개에서 Temporal의 HR@1은 0.7157로 BM25보다 12.11%p 높았다. 이 결과는 반복되는 비교수치와 유사한 문서 구조를 가진 금융공시에서 보고기간 정보가 목표 공시 선택을 보완함을 보여준다.

본 과제(결과물)는 2026년도 교육부 및 대전광역시 재원으로 대전엑스센터의 지원을 받아 수행된 지역성장 인재양성체계(앵커)의 결과입니다.(2026-ANCHOR-06-013)

또한, 본 과제(결과물)는 한남대학교 AI데이터 과학연구소의 지원을 받아 수행된 결과입니다.

참 고 문 헌

- [1] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
- [2] Z. Chen et al., “FinQA: A Dataset of Numerical Reasoning over Financial Data,” *Proceedings of EMNLP 2021*, pp. 3697–3711, ACL, 2021. DOI: 10.18653/v1/2021.emnlp-main.300
- [3] V. Reddy, R. Koncel-Kedziorski, V. D. Lai, M. Krumdick, C. Lovering, and C. Tanner, “DocFinQA: A Long-Context Financial Reasoning Dataset,” *Proceedings of ACL 2024 (Vol. 2: Short Papers)*, pp. 445–458, ACL, 2024. DOI: 10.18653/v1/2024.acl-short.42
- [4] Y. Zhou, F. Zhang, Y. Chen, H. Zhang, P. Nakov, and Z. Xie, “FinCARDS: Card-Based Analyst Reranking for Financial Document Question Answering,” *Findings of ACL 2026*, pp. 24836–24852, ACL, 2026.

- DOI: 10.18653/v1/2026.findings-acl.1244
- [5] Y. Jiang et al., “Fin-RATE: A Real-world Financial Analytics and Tracking Evaluation Benchmark for LLMs on SEC Filings,” Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, pp. 9183–9194, ACM, 2026. DOI: 10.1145/3770855.3817528
- [6] S. Mansouri, D. Saad, M. Wahrenburg, M. Weissel, and F. Woebbecking, “FinRank: An Evidence-Grounded Benchmark for Financial Question Answering and Retrieval over SEC Filings,” arXiv preprint arXiv:2608.07400, 2026. DOI: 10.48550/arXiv.2608.07400
- [7] D. Huwiler, K. Stockinger, and J. Fürst, “VersionRAG: Version-Aware Retrieval-Augmented Generation for Evolving Documents,” arXiv preprint arXiv:2510.08109, 2025. DOI: 10.48550/arXiv.2510.08109
- [8] R. Cymbler, D. Guez, and L. Fabre, “Temporal Misgrounding in Legal RAG: A Versioned-Corpus Benchmark for French Tax Law,” arXiv preprint arXiv:2608.09393, 2026. DOI: 10.48550/arXiv.2608.09393
- [9] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, 3(4), pp. 333–389, 2009. DOI: 10.1561/15000000019
- [10] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of EMNLP-IJCNLP 2019, pp. 3982–3992, ACL, 2019. DOI: 10.18653/v1/D19-1410
- [11] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, “Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,” Proceedings of SIGIR 2009, pp. 758–759, ACM, 2009. DOI: 10.1145/1571941.1572114
- [12] Sentence Transformers, “all-MiniLM-L6-v2,” 2026, Hugging Face Model Card, revision 1110a243fdf4706b3f48f1d95db1a4f5529b4d

41. URL: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2> (accessed August 31, 2026).

저 자 소 개



손정현(Jeongheon Shon)

2021.3 - 현재: 한남대학교 컴퓨터공학과 학사과정
<주관심분야> 빅데이터, 데이터베이스, 자연어처리, 검색 증강 생성(RAG), 대체데이터 분석



장효경(Hyokyung Chang)

1991.2 한남대학교 전자계산공학과 졸업
2012.2 한남대학교 컴퓨터공학과 박사
2021.3-현재: 한남대학교 조교수
<주관심분야> 데이터베이스, 빅데이터, 유비쿼터스 컴퓨팅, 노코드