MECHANISM EVIDENCE CONTRACTS REVERSE BIOMEDICAL RERANKER SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular biologists made 235 of 288 correct adjudications with evidence packets that jointly changed passage order, type labels, and calibrated confidence, versus 203 of 288 with relevance-ranked packets; contradiction misses fell from 28/104 to 16/104. MEC-AUDIT traces this gap upstream by making the evaluator an experimental variable. Eight systems rank common top-100 PubMed Central pools, and their frozen outputs are mapped from one shared adjudication record to entity-level relevance or mechanism support requiring the correct operator, direction, and context. On 918 time-held-out claims and 8,426 judgments, generic judged-set Precision@5 selects BioLinkBERT (0.772), whereas mechanism judged-set Precision@5 selects MEC-RANK (0.742), 0.061 above REACH+BioLinkBERT. Maximizing sets change for 371/918 (40.4%) claims, 26.8% of 25,704 claim–system-pair preferences reverse, and mean rank agreement is 0.48. MEC-RANK realizes the contract with slot interactions, mechanism-preserving hard negatives, separate support and contradiction orders, and calibrated four-way types; contradiction recall is 0.621 versus 0.549 over 596 contradiction-bearing claims. Complete top-100 assessment of 240 claims preserves its leads at 0.711 versus 0.653 Precision@5 and 0.594 versus 0.522 contradiction recall. Across the 4,218-claim, 38,712-pair benchmark, independent labels, transfer, ablations, and the packet study show that evidence semantics can determine both the selected reranker and the scientific decisions it supports.

1 INTRODUCTION

Three passages can all mention metformin and SREBP-1c yet demand three different scientific decisions. Evidence from insulin-resistant HepG2 cells under matched drug exposure can directly support a decrease in SREBP-1c; the same direction in primary hepatocytes or mice transfers only indirectly; an increase under the admissible intervention and context contradicts the claim. Entity-level relevance collapses these cases precisely where mechanism review must separate them.

These three outcomes create a model-selection problem, not just a labeling problem: a leaderboard rewards whichever claim–passage relations its evaluator counts. MEC-AUDIT identifies that choice by intervening after ranking. Eight systems first rank a common top-100 PubMed Central pool; their outputs, one tuple-level adjudication record per observed pair, and the claim-level resampling blocks are then frozen. Projecting those records to either source–target relevance or direct mechanism support changes only the utility map, so each selection response is attributable to evidence semantics conditional on the rankings.

The response is large enough to change comparative conclusions. Generic judged-set Precision@5 selects BioLinkBERT, while mechanism judged-set Precision@5 selects MEC-RANK, ahead of the event-structured REACH+BioLinkBERT comparator by 0.061 (95% CI [0.047,0.076]). Across 918 time-held-out claims, the two evaluators produce different maximizing sets on 371 (40.4%) claims and reverse 26.8% of 25,704 claim–system-pair preferences; mean tie-aware Kendall agreement is 0.48. The disagreement therefore extends from one aggregate winner to the claim and system-pair levels.

MEC-RANK tests what a scorer designed for this target can recover. BioLinkBERT cross-encoding is augmented with entity, operator, polarity, and context interactions, mechanism-preserving hard negatives, an independent contradiction order, four-way evidence typing, and calibrated confidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Complete judging, blinded annotation, overlap pruning, and Reactome transfer test complementary validity conditions; a counterbalanced expert study measures the full packet's decision effect.

Technical contribution and evidence hierarchy. MEC-AUDIT is the paper's central object: a paired evaluator intervention with global, claim-level, pairwise, and rank-agreement responses. The mechanism contract and MEC-RANK make operator, direction, context, adverse evidence, and confidence executable within it. Complete and independent assessment, mechanism-localizing ablations, transfer, and the randomized packet outcome connect selection to system behavior and scientific use. Across 4,218 claims and 38,712 adjudicated pairs, these layers establish that evaluator semantics can change the chosen reranker and the scientific relations presented for expert review.

2 RELATED WORK

Incomplete judgments, construct choice, and ranking utility. IR research studies pooling bias, sparse assessment, and disagreement about a fixed relevance construct (Buckley & Voorhees, 2004; Aslam et al., 2006; Yilmaz & Aslam, 2006; Sakai & Kando, 2008; Sanderson, 2010; Demeester et al., 2016). Measurement modeling identifies the distinct risk that an operational score does not represent the decision construct (Jacobs & Wallach, 2021). MEC-AUDIT makes that mismatch a paired model-selection experiment: one record is mapped through two constructs, and the comparative response replaces an absolute score correction as the target.

Claim verification and biomedical retrieval. FEVER, ERASER, SciFact, and related benchmarks connect claims to supporting or refuting rationales, often coupling retrieval, rationale selection, and veracity (Thorne et al., 2018; DeYoung et al., 2020; Wadden et al., 2020; Wright et al., 2022; Aly et al., 2021; Wadden et al., 2022; Lu et al., 2023). Dense retrieval, cross-encoders, and biomedical pretraining optimize lexical or semantic relatedness (Robertson & Zaragoza, 2009; Karpukhin et al., 2020; Nogueira & Cho, 2019; Khattab & Zaharia, 2020; Beltagy et al., 2019; Lee et al., 2019; Gu et al., 2022; Yasunaga et al., 2022). Our intervention instead changes the definition of useful evidence on frozen within-pool rankings, and its outcome is the selected reranker. Operator, direction, and context make this comparison specific to mechanism evidence rather than generic support/refute classification.

Structured biomedical extraction. PubTator and REACH extract biomedical entities, events, polarity, and context (Wei et al., 2019; Gyori et al., 2017). REACH+BioLinkBERT feeds those fields to a validation-frozen slot scorer. This strongest structured comparator separates access to established event representations from MEC-RANK's joint slot interactions, hard negatives, and dual orders.

Evidence packets and downstream decisions. Retrieval-augmented systems typically evaluate the quality of retrieved context through answer or generation outcomes (Lewis et al., 2020; Izacard & Grave, 2021). Our intervention acts one stage earlier: it asks which reranker should construct the evidence packet when direct support, indirect transfer, and contradiction have different decision roles. The expert crossover evaluates the resulting packet as a joint interface—ordered passages, evidence types, and calibrated confidence—rather than treating topical relevance as a sufficient proxy for adjudication utility.

3 MEC-AUDIT: CONTRACT-CONDITIONED MODEL SELECTION

3.1 TYPED CLAIMS AND EVIDENCE OUTCOMES

A mechanism claim is $q = (s, t, o, p, c)$, with source entity s, target t, intervention type o, expected direction p, and context c (species, tissue or cell line, assay, and condition). A passage maps to $e = (\hat{s}, \hat{t}, \hat{o}, \hat{p}, \hat{c}, h, n)$, adding hedge h and negation n. One shared adjudication record stores these fields plus entity grounding and relation presence. The fixed typed precedence assigns a missing relation to *non-evidence*, an operator or context displacement to *indirect support*, and an admissible relation to *direct support* or *contradiction* by its negation-resolved polarity. Hedge informs confidence. The appendix gives the complete decision sequence.

Table 1: Evidence-contract decisions for a metformin–SREBP-1c mechanism claim.

Item	Contract-matched content
Claim tuple	(metformin, SREBP-1c, pharmacological exposure, −, insulin-resistant HepG2 cells)
Direct support	PMC5952420: metformin lowers SREBP-1c in the matched transformed-cell condition.
Indirect support	PMC3086578: the same direction appears in primary hepatocytes and mice.
Contradiction	Matched metformin exposure and cell context, with an increase in SREBP-1c.
Non-evidence	A candidate discusses metformin-responsive lipid metabolism without measuring SREBP-1c.

Table 1 shows why entity overlap and polarity alone cannot determine evidence utility. PMC5952420 instantiates direct support in the claim-matched transformed-cell setting, whereas PMC3086578 carries the direction into primary hepatocytes and mice and is indirect support. PMC3973184 reports the opposing SREBP-1c direction under atypical PKC co-activation; the operator and context displacement keep this hard candidate outside the contradiction class. A contradiction requires the opposite outcome inside the contract's admissible intervention and context boundary.

3.2 Contract intervention and response signature

Let $\pi_m^S(q)$ be system m's frozen support ordering of the shared candidates and $\pi_m^{S,J}(q)$ its projection onto the design-fixed judged set J_q. The generic evaluator maps $a(q,d)$ to $z^G = g_{st} \in \{0,1\}$ while disregarding operator, direction, and context. The mechanism evaluator maps the same record to $z^T \in \{\text{direct}, \text{indirect}, \text{contradiction}, \text{none}\}$ by the precedence above. Generic Precision@5 uses $u_G(z) = z$; mechanism Precision@5 uses $u_T(z) = \mathbf{1}[z = \text{direct}]$. We call both quantities *judged-set Precision@5* in the primary cohort because the top five are taken after projection onto J_q. The four-way type is richer than this direct-support selection utility: indirect support, contradiction, and non-evidence remain distinct for operator/sign diagnostics, contradiction retrieval, calibration, and packet presentation.

For contract $C \in \{G, T\}$, define the per-claim selection value and its aggregate as

$$V_C(m,q) = \frac{1}{5}\sum_{r=1}^{5} u_C\left(z^C_{q,\pi_m^{S,J}(q,r)}\right), \qquad \bar{V}_C(m) = \frac{1}{|Q|}\sum_{q\in Q} V_C(m,q).$$

The global and claim-level maximizing sets are $S(C) = \arg\max_m \bar{V}_C(m)$ and $s_C(q) = \arg\max_m V_C(m,q)$. Set-valued maximizers retain genuine Precision@5 ties without arbitrary system-specific tie breaking. The claim-level *maximizing-set change* is $|Q|^{-1}\sum_q \mathbf{1}[s_G(q) \neq s_T(q)]$; it records any set inequality and is not described as a unique-winner switch. For system pair (m,n), let $\Delta_C^{mn}(q) = V_C(m,q) - V_C(n,q)$; the pairwise reversal rate averages $\mathbf{1}[\Delta_G^{mn}(q)\Delta_T^{mn}(q) < 0]$ over claims and all system pairs, with either-contract ties contributing zero. For each claim, Kendall's τ_b compares the two eight-system orderings induced by $V_G(\cdot,q)$ and $V_T(\cdot,q)$ while accounting for tied scores; we average the per-claim values. The global winner transition, set change, strict pairwise reversal, and rank agreement form the contract-response signature.

The intervention changes only the evaluator function C: no system is retrained, reranked, recalibrated, thresholded again, or retuned. Precision@5, operator agreement, and sign consistency use $\pi_m^{S,J}$. Operator agreement and sign consistency are the per-claim fractions among the first five projected judged passages whose adjudicated operator and direction, respectively, agree with the claim; both are then macro-averaged over claims. Contradiction recall uses each system's frozen contradiction-score order π_m^C, original top-100 positions, and judged contradictions as its denominator; single-score baselines reuse their support order. A judged contradiction is recovered when its rank under π_m^C is at most 20. Calibration uses every judged pair. Unjudged passages enter none of these numerators or denominators. Claim-blocked intervals reuse identical resamples across evaluators and methods (Guo et al., 2017; Naeini et al., 2015; Nixon et al., 2020; Kumar et al., 2019).

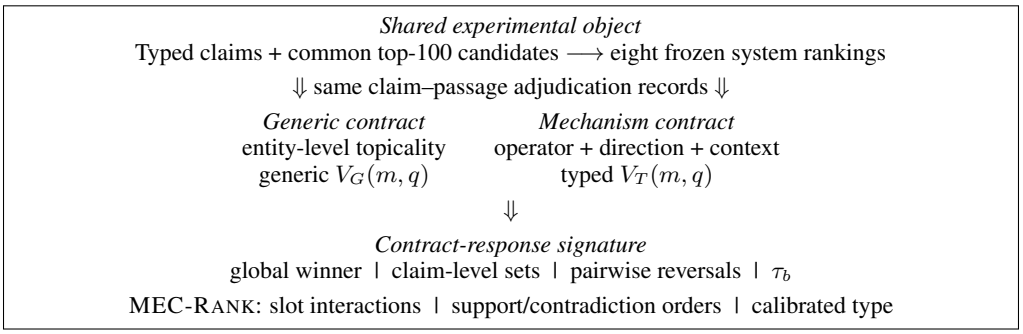


Figure 1: MEC-AUDIT maps shared adjudication records through two evaluators on identical rankings, then measures the selection response. MEC-RANK supplies one frozen output and the typed packet interface.

4 MEC-RANK: CONTRACT-AWARE EVIDENCE RANKING

Structured representation. A PubMedBERT sequence labeler parses claim and passage slots (Gu et al., 2022). BioLinkBERT-large encodes their joint text (Yasunaga et al., 2022); slot projections $u_j(q)$ and $v_j(d)$ for $j \in \{s, t, o, p, c\}$ contribute exact-match, absolute-difference, elementwise-product, and bilinear interactions. The direct-support ranking score is

$$r_\theta(q, d) = w_g^\top g(q, d) + \sum_j w_j^\top [u_j; v_j; |u_j - v_j|; u_j \odot v_j] + u_o^\top A_{op} v_p + u_c^\top A_c v_c,$$

where g is the cross-encoder state. The dense term models the full claim–passage relation, while the slot path exposes the operator, polarity, and context compatibilities on which the contract depends.

Contract-aligned outputs and training. A four-way softmax $P_\theta(y \mid q, d)$ predicts the mutually exclusive evidence type and supplies the calibrated type distribution. A separate contradiction head combines sign reversal, negation, hedge, and context compatibility, allowing adverse evidence to be ordered independently of direct support. Training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda_y \mathcal{L}_{\text{CE}} + \lambda_c \mathcal{L}_{\text{contr}} + \lambda_b \mathcal{L}_{\text{Brier}}.$$

Here $\mathcal{L}_{\text{rank}}$ trains the direct-support order against lower-utility typed outcomes, $\mathcal{L}_{\text{CE}}$ supervises the four mutually exclusive evidence types, $\mathcal{L}_{\text{contr}}$ trains the adverse-evidence order used for recall@20, and $\mathcal{L}_{\text{Brier}}$ regularizes the complete type distribution. The two orders are complementary: contradiction examples must be suppressed in a support packet and promoted in the adverse-evidence list. Mechanism-preserving hard negatives keep the entities fixed while swapping operators, flipping polarity, or changing species, cell line, assay, or negation scope; they target candidates that remain topically compelling while violating one contract slot. Validation selects the loss weights and contradiction threshold, and one temperature fitted to validation logits calibrates the four-way distribution.

Frozen inference interface. At inference, r_θ defines the support order, the contradiction head defines the adverse-evidence order, and P_θ attaches an evidence type and confidence. All three outputs are frozen before the paired evaluator intervention. This separation lets MEC-AUDIT accept rankings from any scorer while testing whether explicit contract structure produces evidence packets that remain useful beyond aggregate ranking scores.

5 EXPERIMENTAL DESIGN

5.1 DATA, ANNOTATION, AND SPLITS

Claims are derived from CTD and KEGG and grounded to PubMed Central passages (Davis et al., 2023; Kanehisa et al., 2021). The benchmark contains 4,218 claims and 38,712 adjudicated pairs (Table 2). Each claim-specific judged set is fixed before evaluation and shared by every system and evaluator. Two Ph.D.-level molecular-biologist authors independently labeled a 500-pair rubric pilot

Table 2: Benchmark accounting, adjudication depth, and metric units.

Item	Value
Mechanism claims	4,218 total = 2,814 CTD + 1,404 KEGG
First-stage retrieval depth	100 passages per claim
Primary paired intervention	918 claims; 8,426 shared judgments
Primary judged set	5–15 passages per claim (mean 9.18)
Benchmark-wide primary judged layer	38,712 query–passage pairs
Benchmark-wide label mix on judged pairs	31.2% direct / 24.7% indirect / 18.4% contradiction / 25.7% non-evidence
Contradiction-recall cohort	596 claims containing a judged contradiction
Additional complete-assessment cohort	240 claims $\times$ 100 passages = 24,000 judgments
Top-k unit	Projected judged rank for support; original pool rank for contradiction
Resource-held-out meaning	CTD train / KEGG test; joint source and schema shift

($\kappa = 0.81$) and jointly adjudicated the remaining retained pairs after the rubric was frozen. For each pair, the record stores source–target compatibility, relation presence, operator, direction, context, hedge, and negation; the generic and mechanism labels are then derived from these fields rather than elicited as unrelated judgments.

Independent validity is measured on 1,200 stratified passages labeled by three blinded Ph.D.-level molecular biologists with model identities and scores hidden. The panel first produces a pre-adjudication decision by majority, with a predeclared rule for three-way ties; a separate adjudication pass establishes the final consensus used for evaluation. Reporting both decisions exposes the exact effect of adjudication rather than folding it into an agreement score.

The random split is 70/10/20. The primary time-held-out test contains post-2022 literature. Resource-held-out trains on CTD and tests on KEGG, jointly shifting source and schema. A second regime trains without Reactome-derived claims (Milacic et al., 2024) and tests 612 claims (5,641 judged pairs). Article IDs, normalized claim tuples, aliases, near-duplicate text, and directed and unordered entity pairs are canonicalized before cumulative overlap pruning.

The primary contract-response estimand is conditional on this design-fixed judged layer: it measures the change in model selection when the evaluator changes on exactly the same observed claim–passage pairs. The projection preserves each system's relative order among observed pairs and assigns no utility to unjudged passages. The complete-assessment cohort removes projection from the typed comparison by adjudicating every top-100 candidate, thereby testing whether MEC-RANK's advantage survives exhaustive assessment of a stratified claim cohort.

5.2 COMPARATORS AND FROZEN-OUTPUT PROTOCOL

The eight-system roster spans lexical retrieval (BM25 (Robertson & Zaragoza, 2009)), biomedical representation learning (PubMedBERT bi-encoder and BioLinkBERT cross-encoder), scientific verification (SciFact), graph assembly (claim-local GraphRAG (Edge et al., 2024)), rubric-prompted classification (GPT-4o (OpenAI, 2024)), event-structured ranking (REACH+BioLinkBERT), and MEC-RANK. This progression separates gains from generic semantic strength, task-specific verification, graph context, prompted reasoning, and explicit event structure.

A single first-stage retrieval run produces the shared top-100 PubMed Central pool before any reranker is scored. GraphRAG builds claim-local graphs only within that pool. GPT-4o receives one typed claim, the rubric, and one passage with deterministic decoding, then emits class probabilities without explicit slot extraction. REACH+BioLinkBERT extracts regulation events, polarity, negation, species, and cell context and feeds them to a validation-frozen slot scorer trained on the same labels and pools. These outputs and the corresponding MEC-RANK outputs are frozen once, then reused under both evaluator contracts.

Table 3: Paired evaluator comparison across eight systems with every time-held-out ranking frozen (918 claims; 8,426 shared judgments). Both columns are projected judged-set Precision@5; higher is better.

System	Generic P@5	Rank	Typed P@5	Rank
BM25	0.613	8	0.421	8
PubMedBERT bi-enc.	0.701	7	0.534	7
BioLinkBERT CE	0.772	1	0.618	3
SciFact	0.728	6	0.587	6
GraphRAG	0.746	5	0.602	5
GPT-4o	0.758	2	0.611	4
REACH+BioLinkBERT	0.748	4	0.681	2
MEC-RANK	0.752	3	0.742	1

Trainable systems use five seeds and are fitted once before contract switching. Hyperparameters are selected on mechanism validation Precision@5 with contradiction recall as the tie-breaker; this common validation rule defines the frozen roster on which evaluator sensitivity is measured, and test labels never enter selection or temperature fitting. Exact aggregate score ties receive average rank, per-claim maximizers remain sets, and threshold ties are assigned to the lower-confidence class. MEC-RANK uses PyTorch 2.3.1, CUDA 12.4, DeepSpeed ZeRO-2, and 8×A100 80GB GPUs; training uses AdamW (Loshchilov & Hutter, 2019), batch size 64, 15 epochs, 1,000 warmup steps, cosine decay, and a learning rate chosen from $\{1, 2, 5\} \times 10^{-5}$. The calibration set contains 4,136 pairs.

Precision@5, operator agreement, and sign consistency average per-claim values over 918 time-held-out claims. Each system's original top-100 order is projected onto the same claim-specific judged set, so the paired estimand compares utility on common adjudicated support without assigning labels to unjudged passages. Contradiction recall measures recovery within the original top 20 over 596 claims containing at least one judged contradiction. ECE uses all 8,426 paired judgments. Their typed-label counts are 2,648 direct, 2,081 indirect, 1,574 contradiction, and 2,123 non-evidence. Claim-blocked bootstrap intervals use identical resamples across contracts and methods.

6 RESULTS

6.1 THE EVALUATOR REVERSES GLOBAL SELECTION ON IDENTICAL RANKINGS

Table 3 gives the decisive controlled result, and Figure 2 visualizes the rank crossing. Generic relevance selects BioLinkBERT by 0.014 P@5 over GPT-4o (95% CI [0.006,0.023]); mechanism evidence selects MEC-RANK by 0.061 over REACH+BioLinkBERT (95% CI [0.047,0.076]). The maximizing sets differ for 371/918 (40.4%) claims. Among all $918 \times \binom{8}{2} = 25{,}704$ claim–system-pair comparisons, 26.8% are strict preference reversals; either-contract ties contribute zero. Mean Kendall's τ_b is 0.48 (95% CI [0.44,0.52]). Thus the aggregate reversal is one visible consequence of a broader reordering, not a close-score anomaly between two systems.

6.2 MEC-RANK LEADS EVERY MECHANISM-EVIDENCE ENDPOINT

Explicit event structure accounts for much of the mechanism-evidence gain: REACH+BioLinkBERT is the runner-up on all five endpoints. MEC-RANK still retains margins of 0.061 P@5 (95% CI [0.047,0.076]) and 0.072 contradiction recall (95% CI [0.052,0.093]) over this comparator. Five-seed means and variability for both systems appear in Table 13.

The generic baselines separate the two behaviors that a scientific packet needs simultaneously. BioLinkBERT leads topical support ranking, whereas SciFact is the strongest generic contradiction retriever. Against the strongest generic comparator for each column, MEC-RANK's difference intervals are [0.109,0.140], [0.113,0.143], [0.077,0.104], [0.147,0.191], and [0.034,0.050], respectively. The support/contradiction leader split is the baseline-only signature of the construct mismatch; MEC-RANK's dual orders close both gaps in one packet.

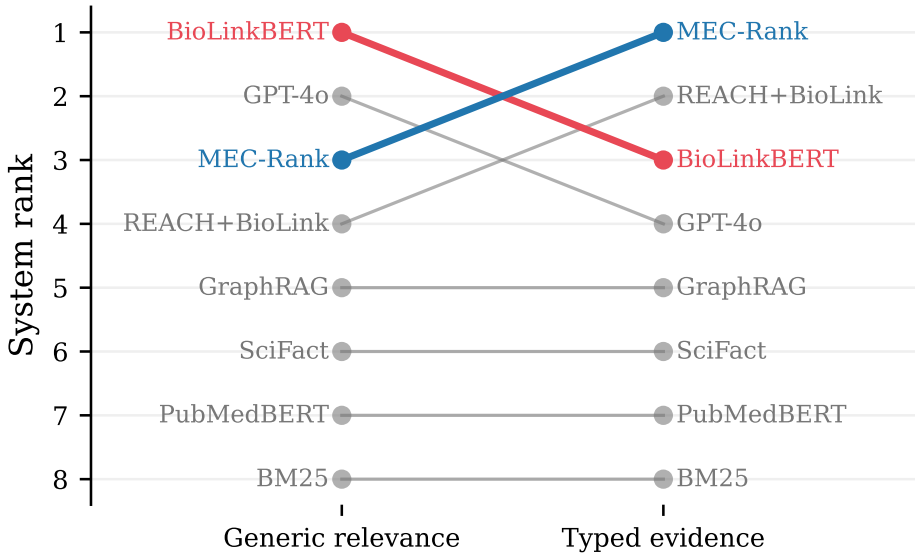


Figure 2: Aggregate ranks after re-evaluating all eight systems' unchanged passage orders under generic and typed contracts. The crossing at rank 1 shows the BioLinkBERT-to-MEC-RANK selection reversal; lower rank is better.

Table 4: Time-held-out mechanism-evidence results on the shared top-100 pool. P@5, operator agreement, and sign consistency use projected judged ranks; contradiction recall uses original pool rank. ECE is lower-is-better.

Method	P@5	Op.	Sign	Contr.	ECE
BM25	0.421	0.483	0.512	0.217	0.182
PubMedBERT bi-enc.	0.534	0.612	0.641	0.341	0.124
BioLinkBERT CE	0.618	0.703	0.724	0.398	0.089
SciFact	0.587	0.671	0.693	0.452	0.103
GraphRAG	0.602	0.694	0.711	0.437	0.097
GPT-4o	0.611	0.698	0.718	0.441	0.092
REACH+BioLinkBERT	0.681	0.785	0.769	0.549	0.073
MEC-RANK	0.742	0.831	0.814	0.621	0.047

6.3 COMPLETE AND INDEPENDENT ASSESSMENT PRESERVES THE TYPED LEADS

We fully adjudicate all 100 candidates for a source-, year-, label-, and mismatch-stratified sample of 240 time-held-out claims, yielding 24,000 judgments. This cohort eliminates rank projection from the mechanism-evidence comparison. MEC-RANK reaches 0.711 P@5, compared with 0.653 for REACH+BioLinkBERT and 0.597 for BioLinkBERT; the difference from BioLinkBERT is 0.114 (95% CI [0.087,0.142]). Contradiction recall is 0.594, compared with 0.522 for REACH+BioLinkBERT and 0.439 for SciFact; the difference from SciFact is 0.155 (95% CI [0.118,0.193]). The all-100 result preserves the same typed winner and structured runner-up while directly evaluating every candidate; all eight rows appear in Table 8.

Independent annotations achieve overall Krippendorff's $\alpha = 0.78$ (95% CI [0.75,0.81]); operator, sign, and context agreement are 0.83, 0.87, and 0.74. The final adjudication changes 185/1200 (15.4%) pre-adjudication panel decisions, and evaluation on that consensus preserves MEC-RANK's advantages over the endpoint-specific generic baselines (Table 5). The overlap flag sets contain 7/918 (0.8%) article, 22/918 (2.4%) near-duplicate-claim, and 85/918 (9.3%) entity-pair matches. These categories overlap and are not additive; removing their union leaves 832 claims and preserves both mechanism Precision@5 and contradiction-recall advantages.

Table 5: Mechanism-evidence validation under three assessment controls. Each cell is MEC-RANK/the strongest generic baseline for that endpoint (BioLinkBERT for P@5; SciFact for contradiction recall).

Evaluation	P@5	Contradiction recall
Complete top-100 judging	0.711/0.597	0.594/0.439
External consensus labels	0.731/0.612	0.603/0.448
Overlap-pruned test	0.728/0.609	0.607/0.445

Table 6: Counterbalanced expert adjudication with relevance-ranked and typed evidence packets. Accuracy and contradiction misses include raw counts.

Endpoint	Relevance packet	Typed packet
Accuracy	203/288 (0.705)	235/288 (0.816)
Median time (minutes)	9.6	7.8
Contradiction misses	28/104 (0.269)	16/104 (0.154)
Confidence Brier	0.213	0.158

6.4 TRANSFER AND ABLATIONS LOCALIZE THE MECHANISM GAINS

On Reactome-held-out transfer, MEC-RANK obtains 0.681 P@5, 0.556 contradiction recall, and 0.068 ECE; REACH+BioLinkBERT obtains 0.639, 0.498, and 0.081, respectively. This second source/schema regime is harder than CTD/KEGG and contains more pathway-composition and context-granularity errors, while preserving the typed advantage. The original resource-held-out result likewise reaches 0.697 P@5 and 0.584 contradiction recall.

The ablations localize the two largest mechanism effects (Table 7 and Figure 3). Removing the contradiction head reduces contradiction recall from 0.621 to 0.412, while removing context fields reduces sign consistency from 0.814 to 0.698. Mechanism-preserving hard negatives and joint training improve every reported axis. In the stress tests (Figure 4), MEC-RANK/BioLinkBERT P@5 is 0.691/0.601 for entity aliasing, 0.634/0.574 for species transfer, 0.612/0.563 for modality mismatch, and 0.658/0.542 for negation/hedge. The $n = 127$ error taxonomy in Table 16 assigns one dominant mechanism to each analyzed resource-held-out failure.

6.5 THE COMPLETE TYPED PACKET IMPROVES EXPERT DECISIONS

The randomized treatment is the complete packet interface: passage order, evidence-type display, and calibrated confidence change together, while source universe, claim, packet size, and source text remain fixed. Participants return a claim-level adjudication and confidence; accuracy uses the study's reference adjudication, and a contradiction miss records failure to identify an adverse-evidence case. The counterbalanced within-subject crossover assigns 24 molecular biologists to 144 claims, producing 576 decisions and 288 observations per condition. Assignment is balanced over participant, claim, condition order, and mechanism family, and participant- and claim-blocked intervals retain both crossed sources of variation. The estimand is the joint effect of the typed packet; contradiction misses are evaluated over the same 104 contradiction-bearing decision opportunities in each condition.

Typed packets improve accuracy by 0.111 (95% CI [0.068,0.154]; odds ratio 1.86, [1.35,2.55]), reduce median time by 1.8 minutes (95% CI [1.2,2.4]; standardized effect 0.42, [0.27,0.58]), and reduce contradiction misses by 0.115 (95% CI [0.031,0.202]; risk ratio 0.57, [0.36,0.91]). The accompanying Brier reduction is 0.055 (95% CI [0.028,0.083]; standardized effect 0.39, [0.19,0.59]). Thus the complete typed-packet condition is more accurate, faster, and more contradiction-aware than the relevance-packet condition.

7 DISCUSSION

The evaluator changes ranking geometry rather than applying a uniform penalty for stricter labels. BioLinkBERT moves from first under topical relevance to third under mechanism evidence, MEC-RANK moves from third to first, 40.4% of claim-level maximizing sets change, and more than one quarter of claim–system-pair preferences reverse. Because the candidate pools, passage orders, shared adjudication records, and resampling units are identical, MEC-AUDIT assigns this within-output disagreement to the evaluator function.

The system comparisons explain why the geometry changes. Generic systems split the desired behaviors: BioLinkBERT is strongest for topical support, whereas SciFact is strongest for contradiction retrieval. Event structure closes much of this gap, and MEC-RANK's slot interactions, mechanism-preserving negatives, and separate contradiction head close it further. Complete top-100 judging preserves the typed winner without projection; blinded consensus labels, overlap pruning, and Reactome transfer preserve the endpoint advantages under different validity checks. The ablations then connect contradiction supervision and context representation to the outcomes they are designed to change.

For scientific RAG and literature triage (Lewis et al., 2020; Izacard & Grave, 2021), the distinction changes the material presented for review. Entity-matched passages can remain fluent and topically coherent while importing a different perturbation, biological setting, or sign. Separating direct, indirect, contradictory, and non-evidence outcomes lets packet construction expose these differences instead of collapsing them into one relevance score. The expert crossover measures the combined interface consequence: typed ordering, labels, and confidence jointly improve accuracy, time, contradiction detection, and calibration.

The response-signature magnitude is conditional on the design-fixed judged layer, eight-system roster, and shared tuning rule; complete judging establishes the typed ordering but does not duplicate every response statistic. Author-adjudicated labels are checked against a blinded challenge set, and the expert experiment identifies the complete packet rather than its interface components. Across the controlled reversal, exhaustive assessment, independent labels, ablations, transfer, and expert outcomes, the evaluated regimes consistently show that mechanism-typed utility selects and serves evidence differently from topical relevance.

8 CONCLUSION

Changing no passage order, a mechanism evidence evaluator selects a different biomedical reranker than entity-level relevance and reverses more than one quarter of claim–system-pair preferences. MEC-AUDIT makes that selection response measurable; MEC-RANK turns operator, direction, and context compatibility into separate support and contradiction orders with calibrated evidence types. Complete assessment preserves the typed ranking, and the resulting packet condition produces 32 more correct expert decisions and 12 fewer contradiction misses across matched 288-decision conditions. Evidence contracts therefore act at two consequential points: they determine which reranker wins and which scientific relations reach expert review.

ETHICS STATEMENT

The expert-adjudication protocol received institutional review approval, and all participants provided informed consent. Public biomedical literature was evaluated without patient-level or private data.

AI USE STATEMENT

Generative AI tools were used to edit the manuscript's language for clarity, readability, and presentation. They were not used to generate data; develop the study's ideas, hypotheses, theory, methods, experiments, mathematical claims, or proofs; implement methods; translate or process data; conduct qualitative analysis; or interpret results. The authors reviewed all AI-assisted edits and take responsibility for the scientific content and final manuscript.

REFERENCES

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pp. 1–13, 2021. URL https://aclanthology.org/2021.fever-1.1/.

Javed A. Aslam, Virgil Pavlu, and Emine Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 541–548, 2006. doi: 10.1145/1148170.1148263.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3615–3620, 2019. URL https://aclanthology.org/D19-1371/.

Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32, 2004. doi: 10.1145/1008992.1009000.

Allan Peter Davis, Thomas C. Wiegers, Robin J. Johnson, Daniela Sciaky, Jolene Wiegers, and Carolyn J. Mattingly. Comparative toxicogenomics database (CTD): update 2023. *Nucleic Acids Research*, 51(D1):D1257–D1262, 2023. doi: 10.1093/nar/gkac833. URL https://doi.org/10.1093/nar/gkac833.

Thomas Demeester, Robin Aly, Djoerd Hiemstra, Dong Nguyen, and Chris Develder. Predicting relevance based on assessor disagreement: analysis and practical applications for search evaluation. *Information Retrieval*, 19(3):284–312, 2016. doi: 10.1007/s10791-015-9275-x.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL*, pp. 4443–4458, 2020. URL https://aclanthology.org/2020.acl-main.408/.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. doi: 10.1145/3458754.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *Proceedings of Machine Learning Research*, 70:1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11):954, 2017. doi: 10.15252/msb.20177651.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*, pp. 874–880, 2021. URL https://aclanthology.org/2021.eacl-main.74/.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2021. doi: 10.1093/nar/gkaa970. URL https://doi.org/10.1093/nar/gkaa970.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *SIGIR*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3792–3803, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. doi: 10.1093/bioinformatics/btz682.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *EMNLP*, pp. 7787–7813, 2023. URL https://aclanthology.org/2023.emnlp-main.483/.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1): D672–D678, 2024. doi: 10.1093/nar/gkad1025.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015. doi: 10.1609/aaai.v29i1.9602. URL https://doi.org/10.1609/aaai.v29i1.9602.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2020. URL https://arxiv.org/abs/1904.01685.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. URL https://arxiv.org/abs/1901.04085.

OpenAI. GPT-4o System Card, 2024. URL https://cdn.openai.com/gpt-4o-system-card.pdf.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Tetsuya Sakai and Noriko Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11(5):447–470, 2008. doi: 10.1007/s10791-008-9059-7.

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. doi: 10.1561/1500000009. URL https://doi.org/10.1561/1500000009.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL*, pp. 809–819, 2018. URL https://aclanthology.org/N18-1074/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4719–4734, 2022. URL https://aclanthology.org/2022.findings-emnlp.347/.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593, 2019. doi: 10.1093/nar/gkz389.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2448–2460, 2022. doi: 10.18653/v1/2022.acl-long.175. URL https://aclanthology.org/2022.acl-long.175/.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8003–8016, 2022. doi: 10.18653/v1/2022.acl-long.551. URL https://aclanthology.org/2022.acl-long.551/.

Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 102–111, 2006. doi: 10.1145/1183614.1183633.

A Additional Results

The supplementary analyses connect component effects, exhaustive assessment, transfer behavior, annotation agreement, seed variation, and calibration to the metric units used in the main text.

Table 7: Time-held-out ablations for MEC-Rank on the judged subset.

Variant	P@5	Op.	Sign	Contr.	ECE
Full model	0.742	0.831	0.814	0.621	0.047
− slot scoring	0.701	0.724	0.731	0.583	0.062
− contradiction head	0.718	0.807	0.791	0.412	0.071
− hard negatives	0.694	0.768	0.752	0.567	0.058
− context fields	0.712	0.741	0.698	0.594	0.053
− joint training	0.687	0.752	0.739	0.541	0.069

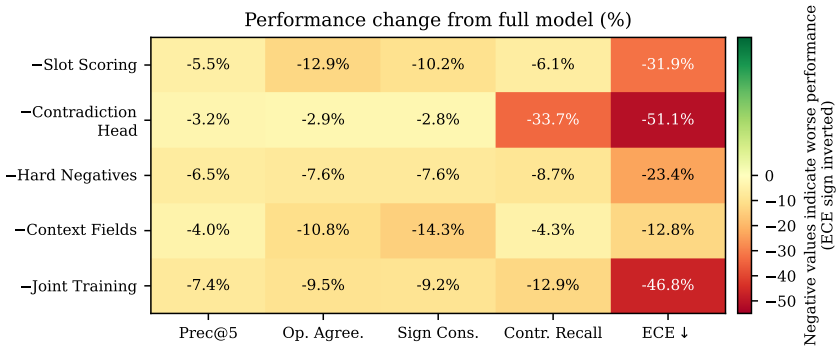


Figure 3: Performance change from the full MEC-Rank model. Negative values indicate worse performance; ECE differences are sign-inverted because lower calibration error is better.

Table 8: Complete top-100 judging on the 240-claim dense cohort.

Method	P@5	Op.	Sign	Contr.	ECE
BM25	0.398	0.462	0.491	0.201	0.191
PubMedBERT bi-enc.	0.506	0.587	0.616	0.319	0.133
BioLinkBERT CE	0.597	0.681	0.702	0.381	0.097
SciFact	0.566	0.648	0.671	0.439	0.111
GraphRAG	0.579	0.671	0.688	0.421	0.105
GPT-4o	0.588	0.676	0.697	0.426	0.100
REACH+BioLinkBERT	0.653	0.756	0.741	0.522	0.079
MEC-Rank	0.711	0.802	0.789	0.594	0.054

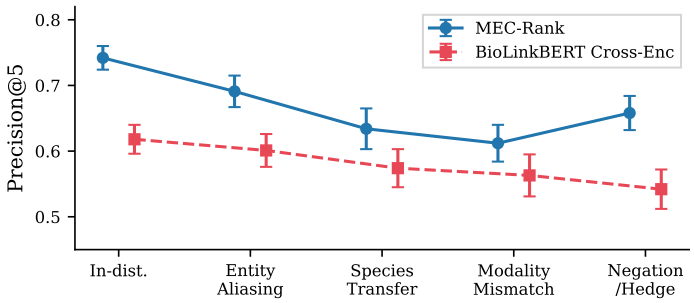


Figure 4: Stress-test Precision@5 for MEC-RANK and BioLinkBERT. Error bars are claim-blocked 95% confidence intervals.

Table 9: Cross-split judged-set evaluation for MEC-RANK.

Split	P@5	Op.	Sign	Contr.	ECE
Random	0.768	0.854	0.838	0.657	0.039
Time-held-out	0.742	0.831	0.814	0.621	0.047
Resource-held-out	0.697	0.793	0.774	0.584	0.061
Reactome-held-out	0.681	–	–	0.556	0.068

Table 10: Independent annotation agreement on 1,200 passages. Label F1 averages each annotator's predictions against the adjudicated consensus.

Measure	Value	Measure	Value
Overall α	0.78	Direct support F1	0.86
Operator agreement	0.83	Indirect support F1	0.71
Sign agreement	0.87	Contradiction F1	0.82
Context agreement	0.74	Non-evidence F1	0.84

Table 11: External annotation confusion counts: the pre-adjudication panel decision (columns) against adjudicated consensus (rows).

Consensus	Direct	Indirect	Contradiction	Non-evidence
Direct	304	18	8	6
Indirect	30	197	13	28
Contradiction	9	15	255	17
Non-evidence	8	22	11	259

The adjudicated consensus contains 336 direct-support, 268 indirect-support, 296 contradiction, and 300 non-evidence passages. Majority and adjudicated labels differ on 185/1200 (15.4%) passages.

Table 12: Reactome-held-out comparison under the typed contract.

System	P@5	Contradiction recall	ECE
BioLinkBERT CE	0.604	0.421	0.095
SciFact	0.575	0.465	0.108
GPT-4o	0.596	0.453	0.099
REACH+BioLinkBERT	0.639	0.498	0.081
MEC-RANK	0.681	0.556	0.068

Table 13: Five-seed means and standard deviations. Point estimates in the main tables use the frozen primary run.

System	P@5	Op.	Sign
PubMedBERT bi-enc.	0.532±0.004	0.610±0.005	0.639±0.006
BioLinkBERT CE	0.620±0.006	0.704±0.005	0.722±0.006
SciFact	0.586±0.007	0.669±0.006	0.691±0.006
REACH+BioLinkBERT	0.681±0.006	0.785±0.005	0.769±0.006
MEC-RANK	0.740±0.005	0.829±0.004	0.812±0.005

System	Contradiction recall	ECE
PubMedBERT bi-enc.	0.338±0.008	0.126±0.005
BioLinkBERT CE	0.401±0.009	0.091±0.004
SciFact	0.450±0.010	0.105±0.005
REACH+BioLinkBERT	0.549±0.009	0.073±0.004
MEC-RANK	0.619±0.008	0.049±0.003

Across all five fitted seeds, MEC-RANK retains the highest mean P@5, operator agreement, sign consistency, and contradiction recall, together with the lowest ECE. The small standard deviations distinguish the evaluator response of the frozen primary outputs from instability due to a single training seed.

Table 14: Four-way calibration diagnostics on the time-held-out paired judgments.

System	Brier	NLL	Adaptive ECE
BioLinkBERT CE	0.176	0.538	0.094
REACH+BioLinkBERT	0.151	0.487	0.078
MEC-RANK	0.118	0.412	0.051

The Brier, negative log-likelihood, and adaptive-ECE ordering agrees with standard ECE in Table 4. Calibration quality therefore accompanies, rather than trades against, the support and contradiction improvements.

Table 15: Direct-support reliability by confidence decile. Confidence is the mean predicted direct-support probability and rate is the empirical direct-support frequency.

Decile	MEC-RANK		BioLinkBERT CE	
	Confidence	Rate	Confidence	Rate
1	0.09	0.10	0.08	0.11
2	0.18	0.19	0.17	0.19
3	0.28	0.27	0.27	0.25
4	0.38	0.39	0.37	0.33
5	0.48	0.49	0.47	0.41
6	0.58	0.57	0.57	0.49
7	0.68	0.69	0.67	0.58
8	0.78	0.77	0.77	0.67
9	0.88	0.86	0.87	0.75
10	0.96	0.94	0.95	0.82

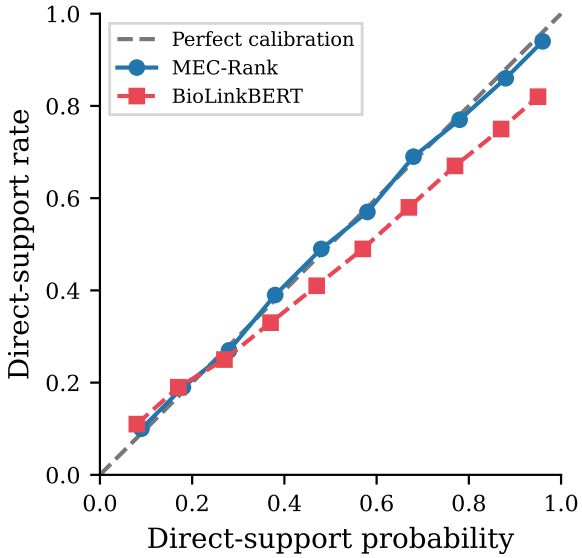


Figure 5: Direct-support reliability on the time-held-out paired judgments. MEC-RANK tracks the empirical rate across confidence deciles more closely than BioLinkBERT.

B Error Analysis

The resource-held-out taxonomy links aggregate transfer errors to the operator, sign, context, candidate-coverage, and annotation mechanisms targeted by the model and protocol.

Table 16: Error taxonomy on resource-held-out failures ($n = 127$). Fractions sum to 99.9% because of rounding.

Failure type	Fraction	Description
Species transfer mismatch	31.5%	Mouse evidence is assigned to a human claim.
Negation scope error	22.8%	Hedge is misclassified as contradiction, or vice versa.
Operator aliasing	18.9%	"Inhibition" and "knockdown" are conflated across contexts.
Few usable judged candidates	15.7%	The pool exists, but only a thin judged slice contains usable evidence.
Annotation convention leak	11.0%	CTD-specific operator conventions transfer poorly to KEGG-derived claims.

C Protocol Details

Shared-record labeling. Annotators first establish source and target grounding and whether the passage asserts a relation between them. A missing relation maps to non-evidence under the mechanism evaluator. For a grounded relation, operator and context are checked before direction: a consequential displacement maps to indirect support, while an admissible pair proceeds to polarity comparison after resolving negation. Aligned polarity maps to direct support and opposed polarity to contradiction. Hedge remains a confidence-bearing field. The generic evaluator reads source–target compatibility from this same record and ignores the remaining fields.

Assessment controls. The complete-assessment cohort is stratified jointly by source, publication year, retained label mix, and mismatch family before all 100 candidates per claim are adjudicated. For external annotation, passages are balanced across the four evidence outcomes and hard operator, sign, context, species, modality, negation, and hedge cases. A predeclared panel rule maps majority decisions and three-way ties to the pre-adjudication label in Table 11; a subsequent adjudication pass produces the consensus. Indirect-support/context confusions dominate. The external cohort is analyzed as a targeted challenge set, while the complete cohort directly removes unjudged candidates from its metric denominator.

Output and decision metrics. The support order supplies judged-set P@5, operator agreement, and sign consistency; the independently frozen adverse-evidence order supplies contradiction recall at original pool depth 20. Calibration is summarized by Brier score, negative log-likelihood, adaptive ECE, and standard ECE on the complete four-way distribution (Table 14). Temperature fitting is performed once on the 4,136 validation pairs and is then frozen for every test regime. For the expert study, assignment is balanced over participant, claim, packet order, and mechanism family; participant- and claim-blocked intervals implement the crossed design for the joint packet treatment.