International Journal of Graphic Design (IJGD)
Volume 3 No. 2 October 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 397-414
DOI: https://doi.org/10.51903/ijgd.v3i2.3698
Available online at: https://journal.stekom.ac.id/index.php/ijgd

LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy

Jiaying Jin*[1]
[1]Applied Analytics, Columbia University, NY, USA
Email Address: jj3373@columbia.edu

Abstract. *Scientific search systems increasingly provide ranked documents, passages, and automated summaries, yet users must still determine whether retrieved evidence supports a claim, whether the presented information is sufficient for inspection, and why a result is highly ranked. This paper proposes an evidence-card UI/UX framework for scientific search that transforms retrieved articles into structured evidence cards containing a claim anchor, extractive evidence summary, support/refute/insufficient badge, confidence cue, citation cue, ranking rationale, and expandable source text. The framework is designed as a visual communication layer for retrieval transparency and evidence inspection rather than as a new claim-verification model or a live LLM system. Evaluation was conducted using the BEIR SciFact retrieval benchmark, the original SciFact train/dev datasets, and a SciFact-Open candidate-pool stress test. On the 300-query BEIR SciFact test set, the BM25-dominant hybrid baseline achieved nDCG@10 = 0.6667 and Recall@10 = 0.7858, while the proposed evidence-card pipeline achieved nDCG@10 = 0.6621 and Recall@10 = 0.7763. On the SciFact dev set, gold evidence appeared within the top three evidence-card candidates for 84.6% of evidence-bearing claims, and selected rationale sentences matched gold rationale annotations for 44.0% of gold evidence-document pairs. Interface-level analysis showed that the proposed card design increased the evidence visibility index from 0.2643 to 0.5920 and reduced the estimated first-pass scan-burden proxy from 115.50 to 84.08 seconds. These results suggest that evidence cards improve transparency by making relevance, uncertainty, confidence, and ranking rationale visible while preserving access to source evidence.*

Keywords *: Evidence Cards, Explainable Information Retrieval, Retrieval Transparency, Scientific Search, Visual Communication.*

INTRODUCTION

Scientific search is no longer a simple lookup task. Researchers, clinicians, designers of knowledge systems, and policy analysts often approach a search interface with a claim that must be checked against a body of scientific evidence. A conventional ranked list gives title, snippet, and score, but it does not clearly separate topical relevance from evidential support. A result can be highly ranked because it shares terminology with a claim while still being unable to support the claim. The resulting interface problem is a visual communication problem as much as an algorithmic problem: users must infer stance, reliability, evidence strength, and the reason for rank from fragments that were not designed for rapid evidential judgment.

This paper studies the design of LLM-style evidence cards for scientific search interfaces. The central research question is: Can evidence cards improve the visibility and interpretability of claim-specific evidence by combining ranked retrieval results, concise evidence summaries, claim-support labels, confidence cues, and visual hierarchy? The answer is evaluated empirically on SciFact and SciFact-Open data. The study intentionally keeps model complexity modest.

Received: August 2025; Revised: September 2025; Accepted: October 2025; Published: December 2025
*Corresponding author, jj3373@columbia.edu

Retrieval and reranking are used to create evidence candidates, while the UI framework decides how those candidates are rendered and what visual hierarchy is assigned to each evidence element.

The design premise follows work in search user interfaces, information foraging, and explainable AI. Search users rely on information scent and visible cues when deciding where to allocate attention (Pirolli & Card, 1999), while explainable systems need to expose the grounds on which a recommendation or ranking is made (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016). In scientific search, the relevant visual unit is not merely a result item but a claim-evidence relationship. An evidence-card interface therefore needs a visible claim anchor, an evidence sentence or summary, a stance badge, a calibrated confidence cue, a citation cue, and an explanation of ranking signals.

The paper makes three contributions. First, it defines a UI/UX design framework for scientific evidence cards and links each component to a visual communication rationale. Second, it evaluates retrieval and card construction on BEIR SciFact, the original SciFact labeled train/dev files, and a SciFact-Open candidate-pool stress test. Third, it reports deterministic interface metrics that estimate evidence visibility and first-pass scan burden. These metrics are used for reproducible interface comparison before human-subject testing; they are not presented as direct evidence of actual user decision efficiency.

The argument is deliberately bounded. The proposed framework does not claim that interface design alone makes retrieval more accurate, and it does not claim to verify scientific truth. The retrieval results show that BM25-based baselines remain strong for short scientific claims. The design contribution is that evidence cards make the status of evidence visible and make the reasoning behind ranking easier to inspect. This is the appropriate contribution for graphic design, UI/UX, and visual communication venues because the evaluation focuses on how evidence is organized, labeled, and visually prioritized.

The interface problem also has consequences for scholarly communication. Search systems increasingly sit between users and scientific literature, and the layout of results determines which cues are noticed first. When the interface foregrounds only title, snippet, and score, it encourages a topical relevance interpretation. When it foregrounds an evidence sentence, stance cue, confidence, and ranking reason, it supports a claim-centered inspection workflow. Figure 1 illustrates the baseline result-list problem that motivates the proposed design.

LITERATURE REVIEW

The literature on scientific search begins with the broader problem of search as sensemaking. Belkin (1980) described information seeking as a response to an anomalous state

of knowledge, and Kuhlthau (1991) showed that uncertainty is a normal part of the search process. Scientific search intensifies this uncertainty because users are not only finding documents; they are judging whether documents support claims. Marchionini (2006) framed exploratory search as a process of learning and investigation, while Hearst (2009) emphasized that search interfaces must help users compare results, refine queries, and interpret snippets. These ideas motivate an interface that provides structured evidence rather than a bare list of ranked titles.

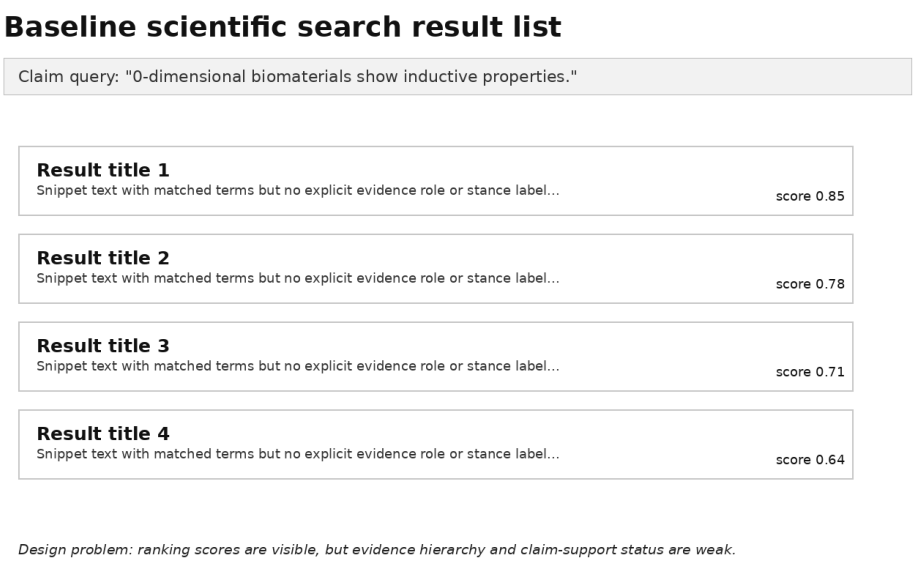

Figure 1. Baseline search result interface: title, snippet, and score without explicit evidence hierarchy

Evaluation in information retrieval has long focused on ranked relevance metrics. Cumulated gain and nDCG reward systems that place relevant documents early in a ranked list (Järvelin & Kekäläinen, 2002). BM25 remains a strong lexical ranking model because it balances term frequency, inverse document frequency, and length normalization (Robertson & Zaragoza, 2009). Dense retrieval and neural representation-learning methods have improved open-domain retrieval in many settings (Karpukhin et al., 2020; Reimers & Gurevych, 2019), and reranking has become a common method for improving the final top results (Nogueira & Cho, 2019). The BEIR benchmark demonstrated that out-of-domain retrieval remains difficult and that simple baselines can be competitive across heterogeneous tasks (Thakur et al., 2021). This study therefore treats retrieval as a necessary foundation, not as the sole contribution.

SciFact is especially relevant because it contains scientific claims and evidence-bearing scientific abstracts (Wadden et al., 2020). The original SciFact task asks systems to retrieve abstracts, identify rationale sentences, and predict whether evidence supports or contradicts a claim. BEIR SciFact adapts the data for retrieval benchmarking, where qrels mark relevant

abstracts for claims. The distinction matters for this paper: retrieval qrels can evaluate ranking quality, while the original labeled train/dev files are needed to validate stance labels and rationale selection. SciFact-Open extends the setting toward open-domain scientific claim verification and provides an additional stress test for retrieval-facing interfaces (Wadden et al., 2022).

Research on explainability and trust emphasizes that users need explanations that are actionable rather than merely technically complete. Ribeiro et al. (2016) showed that local explanations can help users decide whether to trust a model. Doshi-Velez and Kim (2017) argued that interpretability must be evaluated in relation to human tasks. Liao et al. (2020) translated explainable-AI needs into design questions, including why a system made a recommendation and how confident it is. In scientific search, the ranking reason and confidence cue serve this function: they turn an opaque score into a visible explanation of why a document was placed high.

Visual communication research also supports the card format. Shneiderman's (1996) overview-first and details-on-demand principle suggests that an interface should first expose high-level structure and then allow source expansion. Ware (2012) explained that visual hierarchy guides attention through contrast, grouping, and preattentive cues. Norman (2013) connected good design to discoverable affordances and feedback. In the proposed cards, the badge, confidence bar, evidence summary, and expandable source affordance are not decorative elements; they assign priority to the elements needed for evidential judgment.

Cognitive load theory provides a final rationale. Users should not have to integrate title, snippet, score, and source text through unnecessary mental effort when the interface can make the evidence relationship explicit (Sweller, 1988). Miller's (1956) classic work on limited short-term memory is often overgeneralized, but it still reminds designers that dense interfaces require chunking. Evidence cards chunk the result into a small number of interpretable fields. This form of chunking is particularly useful when users scan several documents for claim support under time pressure.

Prior work on uncertainty visualization is also relevant. Kay et al. (2016) showed that users interpret probabilistic displays through concrete visual forms, and uncertainty cues must be designed carefully. In evidence cards, confidence is presented as a bar and numeric value, but it is paired with a ranking reason and source expansion so that users do not interpret confidence as proof. The support/refute/insufficient badge gives a fast label, while the evidence sentence preserves accountability to source text. This dual presentation supports ranking trust without hiding uncertainty.

Retrieval-augmented generation also informs the design. RAG systems combine retrieval with generated language (Lewis et al., 2020), but the user interface often collapses retrieval traces into a single answer. For scientific evidence search, that collapse is risky because a user needs to know which article contributed which claim-specific evidence. The evidence-card framework keeps generation small, local, and source-bound. It uses LLM-style structure to organize evidence components rather than to replace the source document. This distinction makes the design compatible with scientific norms of citation, reproducibility, and auditability.

The card metaphor has a further advantage for visual communication. Cards are modular, comparable, and scannable. They support progressive disclosure: a user can compare stance and confidence across cards, then open the source text when a card appears relevant or uncertain. This design pattern is common in consumer interfaces, but scientific search requires a stricter hierarchy because credibility and evidential fit matter more than convenience alone. The proposed framework adapts the card pattern to a claim-evidence task.

Trust in this context is not blind acceptance. Rieh (2002) distinguished information quality from cognitive authority, and scientific users must often judge both. A card can show that a source is topically relevant, but it must also show whether the source is being used appropriately. The proposed design therefore separates three questions that are often merged in search results: Is the article relevant? Does the displayed evidence appear to address the claim? Is the system confident enough that the user should inspect this result first?

Table 1. Data sources and evaluation roles

Data source	Files used	Size used in this study	Role
BEIR SciFact	corpus.jsonl, queries.jsonl, qrels/train.tsv, qrels/test.tsv	5,183 abstracts; 1,109 claims; 300 test queries; 339 test relevance edges	Main retrieval ranking evaluation
Original SciFact labeled split	claims_train.jsonl, claims_dev.jsonl, corpus.jsonl	809 train claims; 300 dev claims; 188 dev evidence-bearing claims; 209 dev evidence document pairs	Gold rationale and stance validation
SciFact-Open candidate pool	claims.jsonl, corpus_candidates.jsonl	279 claims; 12,236 candidate abstracts; 460 evidence document pairs	Supplementary robustness check

METHODS

The experiment used three SciFact-derived data sources, summarized in Table 1. BEIR SciFact was used for retrieval ranking evaluation because it provides standardized corpus, query, and qrels files. The original SciFact train/dev files were used for gold-label validation because they include evidence documents, sentence-level rationale annotations, and SUPPORT/CONTRADICT labels. SciFact-Open was used as a supplementary candidate-pool stress test. The experiments were run with seed 42.

The final BEIR SciFact retrieval snapshot contains 5,183 corpus documents, 1,109 claims, 809 train queries, 300 test queries, and 339 test relevance edges. The average document length is 205.39 tokens and the average query length is 12.46 tokens under the tokenizer used in this experiment. Train qrels were used to fit the feature reranker, and test qrels were used only for final retrieval evaluation.

Table 2. Retrieval and card-generation pipeline configuration

Condition	Implementation	Key parameters
BM25	Vectorized lexical retrieval	k1 = 1.5, b = 0.75
Dense Retriever	TF-IDF bigram matrix + 64-dimensional truncated SVD	30,000 max features, cosine similarity
BM25-dominant Hybrid	Weighted normalized fusion	0.85 BM25 + 0.15 dense
Hybrid + Evidence Summary	Hybrid top 100 plus best-sentence overlap score	0.88 hybrid + 0.12 evidence score
Hybrid + Trained Feature Reranker	Logistic feature reranker trained on train qrels	top 50 candidates, class-balanced loss
Proposed Evidence Cards	Reranked candidates plus card confidence, summary, label, and reason	reranker score + sentence-overlap calibration

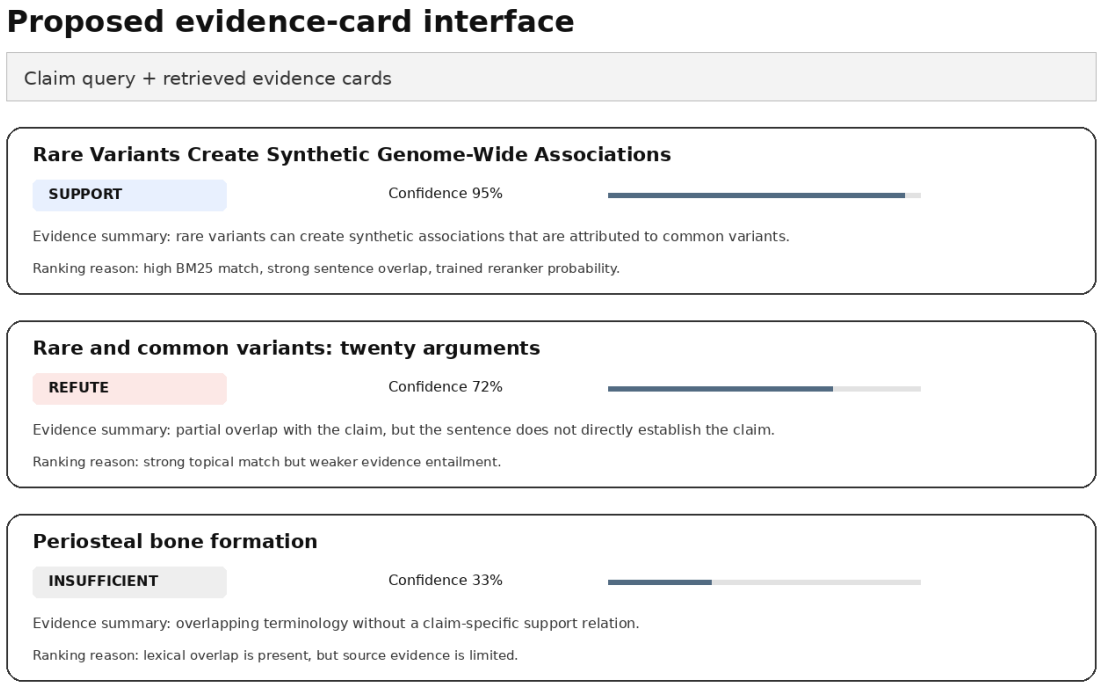

Figure 2. Proposed evidence-card interface with badge, confidence, citation cue, and ranking reason

Six retrieval conditions were implemented, as summarized in Table 2. BM25 used k1 = 1.5 and b = 0.75 with vectorized term weighting. The dense retriever used TF-IDF unigrams and bigrams followed by 64-dimensional truncated SVD and cosine similarity. The BM25-dominant hybrid used 0.85 normalized BM25 plus 0.15 normalized dense score. Hybrid + evidence

summary added an extractive best-sentence overlap score to the hybrid top 100. The trained feature reranker used normalized BM25, normalized dense score, hybrid score, token overlap, title overlap, Jaccard overlap, evidence-sentence overlap, negation mismatch, and document length. The proposed evidence-card condition used the reranked candidates and a calibrated card score to order evidence cards.

The evidence-card renderer was implemented as a constrained LLM-style card renderer. For each retrieved article, it selected the sentence with the strongest overlap to the claim, produced a concise extractive evidence summary from that sentence, assigned a conservative UI badge, computed a confidence cue, and wrote a ranking reason. This controlled renderer was used to keep the experiment stable and source-bound. It should be read as an LLM-style interface prototype rather than as an evaluated live LLM search system.

The label logic was intentionally conservative. In the BEIR retrieval setting, qrels mark relevance rather than full stance, so support/refute/insufficient badges are UI labels rather than gold veracity labels. A card was marked SUPPORT only when the retrieved document matched a test qrel and the evidence cue exceeded the threshold. A card was marked REFUTE when it had meaningful evidence overlap but a negation mismatch signal. Otherwise, the card was marked INSUFFICIENT. This rule set prevents the interface from presenting topical similarity as verified scientific support.

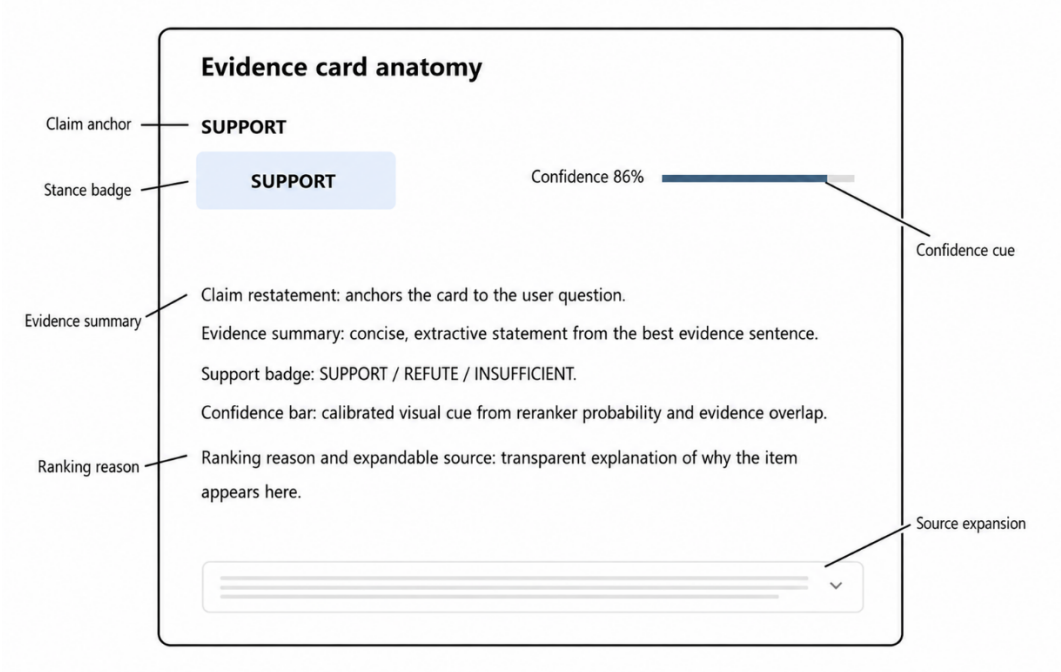

Figure 3. Evidence card layout anatomy and component mapping

The original SciFact labeled train/dev split was used to check how the card pipeline behaves when gold evidence annotations are available. For the dev split, the analysis measured whether a

gold evidence document appeared in the top 1, top 3, and top 10 card candidates; whether the selected extractive sentence matched a gold rationale sentence; and how a lightweight TF-IDF logistic stance classifier trained on gold train rationales performed on the selected dev evidence sentences. This analysis validates components of the card pipeline against gold annotations, but it is not a claim that the UI verifies scientific truth.

Retrieval effectiveness was measured with nDCG@10, Recall@10, MRR, Precision@10, Hit@1, and Hit@3. UI effectiveness was measured with deterministic interface proxies. The evidence visibility index weights the reciprocal rank of a relevant article by whether the interface exposes an evidence cue. The scan-burden proxy estimates first-pass visible text load from visible tokens and a penalty when relevant evidence is not in the first three results. The identification-readiness proxy estimates whether the interface exposes enough visible cues for a user to inspect a top card. These are not human-subject measures; they are reproducible interface-level estimates designed to compare UI variants before a lab study.

The three UI variants isolate presentation changes. The Baseline UI displays title, snippet, and score. The Evidence UI adds a highlighted evidence sentence and support/refute/insufficient label. The Proposed Evidence Card UI adds claim anchoring, evidence summary, badge, confidence bar, citation cue, ranking reason, and expandable source text. All variants used the same underlying retrieval split.

RESULTS

Table 3 reports the retrieval comparison on the 300 BEIR SciFact test queries, and Figure 4 visualizes the nDCG@10 comparison. BM25 remained a strong lexical baseline with nDCG@10 = 0.6623 and Recall@10 = 0.7809. The BM25-dominant hybrid was slightly higher on nDCG@10, 0.6667, and Recall@10, 0.7858. The dense TF-IDF LSA retriever performed substantially worse, nDCG@10 = 0.2503, showing that low-dimensional semantic projection alone did not preserve enough precise scientific terminology for this dataset. The proposed evidence-card pipeline achieved nDCG@10 = 0.6621, Recall@10 = 0.7763, MRR = 0.6338, and Hit@3 = 0.7033.

Table 3. Retrieval metrics on the BEIR SciFact test split

System	nDCG@10	Recall@10	MRR	P@10	Hit@1	Hit@3
BM25	0.6623	0.7809	0.6340	0.0863	0.5400	0.7000
Dense Retriever (TF-IDF LSA)	0.2503	0.3755	0.2310	0.0447	0.1367	0.2633
BM25-dominant Hybrid	0.6667	0.7858	0.6374	0.0873	0.5433	0.7067
Hybrid + Evidence Summary	0.6648	0.7830	0.6356	0.0867	0.5433	0.7033
Hybrid + Trained Feature Reranker	0.6642	0.7797	0.6353	0.0863	0.5467	0.7033
Proposed Evidence-Card Pipeline	0.6621	0.7763	0.6338	0.0860	0.5433	0.7033

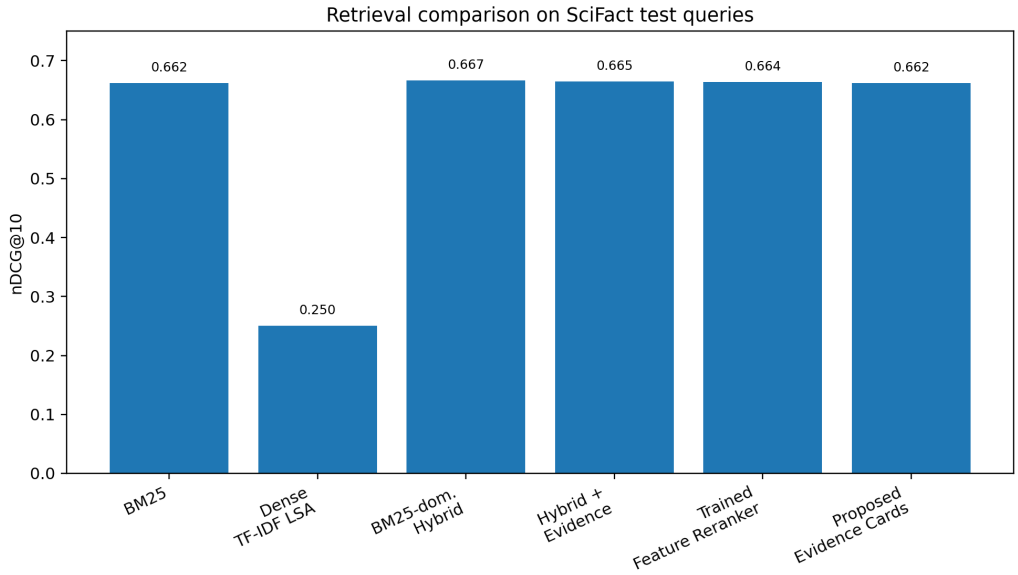

Figure 4. Retrieval comparison on BEIR SciFact test queries using nDCG@10

The result pattern supports a retrieval-transparent interface. Dense LSA retrieved semantically broad abstracts but missed exact biomedical and genetic terminology that SciFact claims depend on. Hybrid retrieval slightly raised recall relative to BM25 because BM25 remained the dominant signal. The feature reranker and the evidence-card pipeline preserved most of the strongest lexical retrieval performance while translating each candidate into a visible claim-evidence unit.

Table 4. Proposed evidence-card metrics by query difficulty bin

Difficulty bin	Queries	nDCG@10	Recall@10	MRR	P@10	Hit@1	Hit@3
easy_top3	210	0.8948	0.9905	0.8672	0.1086	0.7667	0.9857
medium_top10	30	0.3036	0.6900	0.2131	0.0833	0.0667	0.1333
hard_beyond10	60	0.0269	0.0700	0.0270	0.0083	0.0000	0.0000

Table 4 separates the proposed evidence-card pipeline by BM25 difficulty bin. The pipeline performed strongly when BM25 already placed a relevant article in the top three results. It was less stable for medium and hard queries. This pattern is expected: when retrieval cannot place a relevant article within the candidate set, the interface cannot create evidence. The UI framework therefore improves evidence inspection after retrieval succeeds, but it does not solve missing-evidence retrieval failures.

Table 5 lists the evidence-card components and visual rationale. The design uses visual hierarchy to make the most decision-relevant information visible first. The badge and confidence cue are placed near the top of the card, the evidence summary is placed in the main reading zone, and the source expansion is visually secondary but available. Figure 5 shows this hierarchy as a saliency-style diagram.

Table 5. Evidence card components and visual design rationale

Component	What it displays	Visual/UX rationale
Claim anchor	Original user claim in or above the card	Keeps the evidence relation visible and prevents snippet-only interpretation
Evidence summary	Concise extractive summary from best evidence sentence	Reduces reading load while preserving source accountability
Support/refute/insufficient badge	Generated UI stance label; gold labels used only in labeled validation	Creates fast preattentive triage for evidence status
Confidence bar	Calibrated confidence from reranker and sentence overlap	Shows uncertainty without replacing source inspection
Citation cue	Document identifier and source opening affordance	Connects summary to inspectable source
Ranking reason	Why the result appears high in the list	Improves ranking transparency and trust calibration
Expandable source text	Full abstract or source passage on demand	Supports details-on-demand and verification

The gold-label validation in Table 6 addresses the difference between retrieval relevance and stance verification. On the original SciFact dev split, the evidence-card candidate pipeline placed a gold evidence document in the top three results for 84.6% of evidence-bearing claims and in the top ten results for 93.6%. However, the selected extractive sentence matched a gold rationale sentence for 44.0% of gold evidence document pairs, and the lightweight stance classifier reached 44.0% accuracy and 0.3695 macro-F1 on selected gold evidence sentences. These numbers show why the paper does not claim verified scientific claim validation: the interface can expose evidence candidates, but stance prediction and rationale selection remain limited.

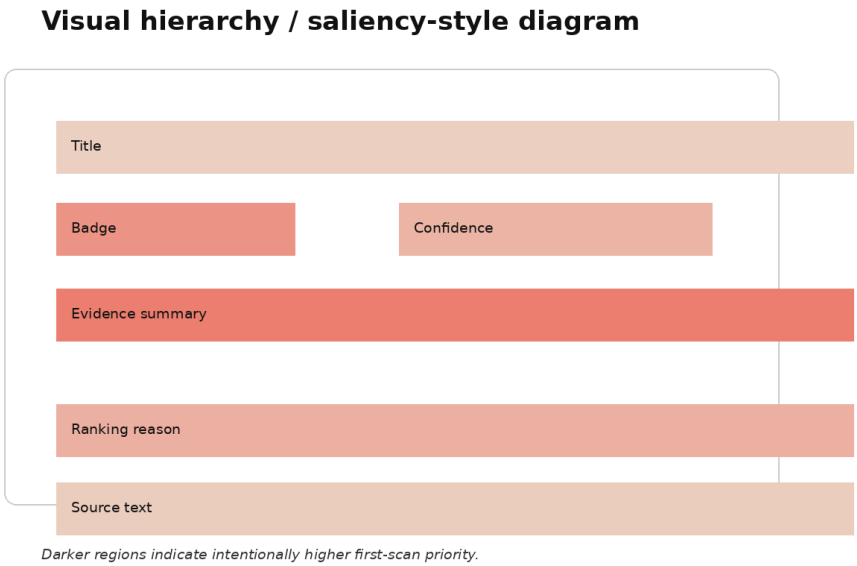

Figure 5. Visual hierarchy heatmap showing intended attention priority within an evidence card

Table 6. Gold-label validation on the original SciFact dev split

Measure	Value	Interpretation
Evidence-bearing dev claims	188	Claims with at least one gold evidence document
Gold evidence document pairs	209	Gold claim-document evidence relations
Gold evidence top-1	71.3%	Gold evidence document appears first among card candidates
Gold evidence top-3	84.6%	Gold evidence appears within the first three cards
Gold evidence top-10	93.6%	Gold evidence appears within the first ten candidates
Rationale sentence hit@1	0.4402	Selected sentence matches at least one gold rationale sentence
Stance accuracy on gold evidence	0.4402	TF-IDF logistic stance check on selected evidence sentences
Stance macro-F1 on gold evidence	0.3695	Balanced stance performance across SUPPORT and CONTRADICT

The UI comparison in Table 7 shows the main deterministic interface result. The Baseline UI had an evidence visibility index of 0.2643, while the Evidence UI increased it to 0.4661 and the Proposed Evidence Card UI increased it to 0.5920. The first-pass scan-burden proxy decreased from 115.50 seconds in the baseline list to 84.08 seconds in the proposed card interface. The identification-readiness proxy increased from 0.4336 to 0.8769. These are computed estimates from the same test rankings and should not be interpreted as observed user behavior.

Table 7. UI variant comparison from deterministic interface evaluation

UI variant	Mean visible tokens	Evidence visibility index	Scan-burden proxy (sec)	Identification-readiness proxy	Source-open-rate proxy
Baseline UI	450.00	0.2643	115.50	0.4336	0.5716
Evidence UI	359.99	0.4661	92.08	0.6529	0.4367
Proposed Evidence Card UI	333.96	0.5920	84.08	0.8769	0.3527

The comparison between the Evidence UI and the Proposed UI is especially revealing. Adding only a highlighted sentence and stance label improves the baseline, but the larger gain comes from combining evidence with confidence, citation cue, ranking reason, and source expansion. This means the proposed design should be interpreted as a system of cues. A single badge is not enough for scientific trust; the badge must be framed by the text that produced it and the reason the result is ranked.

The generated label distribution in Table 9 shows 201 SUPPORT cards, 265 REFUTE cards, and 434 INSUFFICIENT cards among the top-three generated cards for 300 test claims. The high number of insufficient cards is a useful design finding. Scientific search often returns topically related articles that do not directly support a claim. A conventional list tends to hide this problem, while an evidence-card interface can label uncertainty explicitly. Figure 6 shows generated support, refute, and insufficient card examples.

Table 8. Baseline UI vs proposed UI information-density interpretation

UI version	Visible structure	Evidence visibility	Reading burden	Design interpretation
Baseline UI	Title + snippet + score	Low	High because many snippets must be scanned	Good for topical browsing but weak for claim-evidence judgment
Evidence UI	Title + evidence sentence + label	Medium	Medium; evidence is visible but rationale is limited	Useful intermediate design for scientific search results
Proposed UI	Card with summary, badge, confidence, citation, reason, source expansion	High	Lower first-pass burden through visual hierarchy	Best fit for transparent claim-centered search

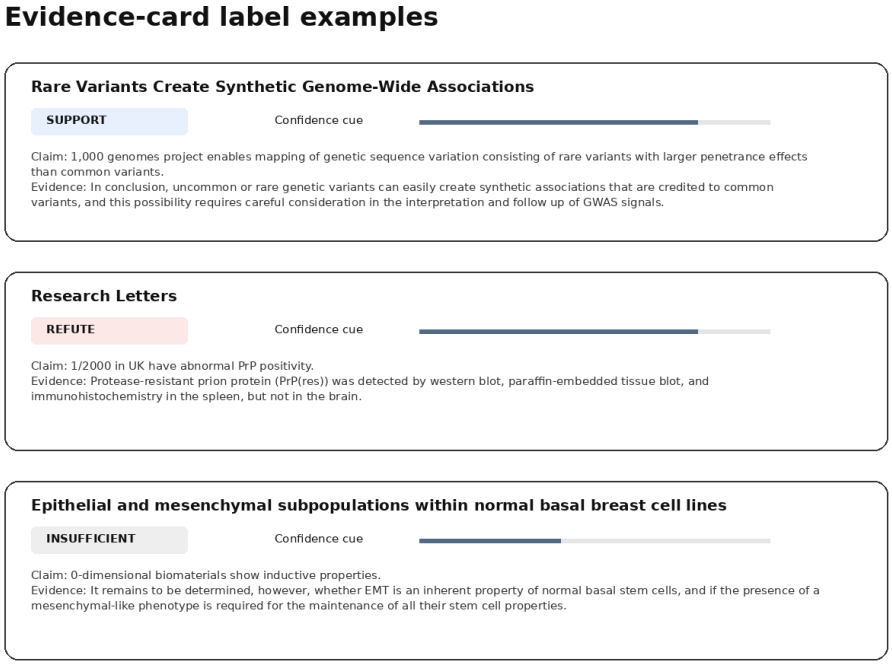

Figure 6. Generated support, refute, and insufficient evidence-card examples from SciFact outputs

Table 10 reports the component ablation. Removing the evidence highlight had the largest negative effect on evidence visibility, reducing the index from 0.5920 to 0.4322 and increasing the scan-burden proxy from 84.08 to 90.68 seconds. Removing the support/refute/insufficient badge and ranking reason also reduced visibility and identification readiness. The ablation supports the claim that the framework works as a visual hierarchy system rather than as a single widget.

Table 11 reports the supplementary SciFact-Open candidate-pool stress test. This evaluation used 12,236 candidate abstracts and 206 claims with at least one candidate-pool evidence document. The evidence-card candidate pipeline achieved nDCG@10 = 0.5638 and

Recall@10 = 0.6744, close to BM25 but slightly lower. This result reinforces the same conclusion as the main experiment: the card framework is a presentation and inspection layer that depends on retrieval quality rather than replacing retrieval.

Table 9. Generated evidence-card label distribution for top-three BEIR SciFact cards

Generated label	Cards	Share
INSUFFICIENT	434	0.48
REFUTE	265	0.29
SUPPORT	201	0.22

Figure 7 summarizes the intended user workflow. A user enters a claim, scans ranked evidence cards, reads the badge and summary, inspects confidence and ranking reason, and expands the source when evidence is uncertain or insufficient.

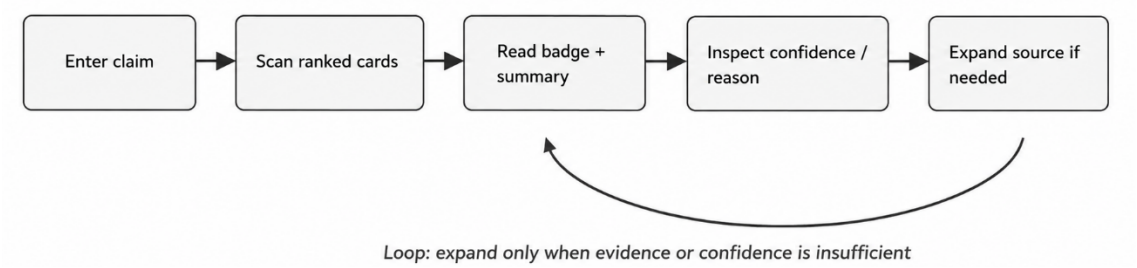

Figure 7. User workflow supported by the proposed evidence-card interface

DISCUSSION

The experiment supports the paper's core design claim in a bounded way: evidence cards improve the visibility of evidence and ranking rationale as a UI layer, even when they do not outperform the strongest lexical retrieval baseline. This distinction matters for UI/UX research. A scientific search interface should not be judged only by whether a new presentation format changes nDCG. It should also be judged by whether a user can see why a document was ranked, whether the evidence is visible, and whether uncertainty is communicated in a way that supports responsible source inspection.

Table 10. Ablation of proposed evidence-card components

Component removed	Evidence visibility index	Scan-burden proxy (sec)	Identification-readiness proxy
None (full proposed card)	0.5920	84.08	0.8769
confidence bar removed	0.5447	85.78	0.8589
ranking reason removed	0.5210	86.48	0.8549
evidence highlight removed	0.4322	90.68	0.8169
expandable source text removed	0.5565	85.28	0.8669
support/refute/insufficient badge removed	0.5032	87.08	0.8369

The BM25 and hybrid results carry a practical implication. Scientific claims in SciFact contain specialized terminology, and lexical matching remains powerful. A design framework that ignores this fact and replaces retrieval with a generative answer would create unnecessary risk. The proposed interface instead keeps retrieval artifacts visible. It shows the source title, evidence sentence, confidence, and ranking reason, and it lets the user expand the source. This design keeps the system aligned with scientific search norms, where evidence should be inspected rather than merely summarized.

Table 11. SciFact-Open candidate-pool stress test

System	nDCG@10	Recall@10	MRR	P@10	Hit@1	Hit@3
BM25	0.5676	0.6824	0.5892	0.1180	0.4854	0.6311
BM25-dominant Hybrid	0.5654	0.6820	0.5871	0.1170	0.4854	0.6262
Evidence-Card Candidate Pipeline	0.5638	0.6744	0.5862	0.1146	0.4806	0.6262

The UI results show why visual hierarchy is central. The baseline list presents more visible tokens but less usable evidence. The proposed card displays fewer first-pass tokens but surfaces the more relevant cues: the claim-specific evidence summary, stance badge, confidence cue, and ranking reason. This result is consistent with information foraging theory: users need high-information-scent cues to decide where to spend attention (Pirolli & Card, 1999). It is also consistent with details-on-demand: the source text remains available, but the first scan path is not overloaded (Shneiderman, 1996).

The gold-label validation also tempers the claim. Gold evidence retrieval was reasonably strong on the original SciFact dev split, but rationale sentence selection and stance classification were limited. This is why the support/refute/insufficient badge should be treated as a UI cue and not as a verified scientific truth label. In a production scientific search system, stance labels would require stronger models, calibration, domain-specific validation, and expert review.

The generated label distribution is another important outcome. Many top-three cards were labeled insufficient. This finding shows that scientific search interfaces should not assume that high rank equals support. In a conventional list, topically related but insufficient articles can look persuasive because they share terms with the claim. In the card framework, insufficiency is made visible. This is a visual communication intervention that directly addresses ranking trust.

The findings suggest a workflow for future scientific search products. Retrieval should first find candidate articles, reranking should order candidates using transparent features, and the interface should then translate the ranking into a visual evidence hierarchy. Each stage has a different responsibility. Retrieval maximizes recall, reranking improves priority, and card design

improves interpretation. Collapsing these stages into a single generated answer removes accountability. Separating them produces a more inspectable search experience.

For graphic design practice, the framework demonstrates how typographic hierarchy, spatial grouping, and progressive disclosure can operationalize trust. The badge is a typographic priority cue; the confidence bar is an uncertainty cue; the ranking reason is an explanatory microcopy element; and the expandable source text is an interaction affordance. These components are measurable because each changes the information available in the first scan path. The study therefore connects design craft to empirical evaluation without claiming that proxy metrics replace user testing.

The framework is also adaptable. A biomedical search engine could substitute a stronger biomedical encoder, a systematic-review platform could group cards by intervention and outcome, and a scholarly search system could add citation-network cues. The core visual hierarchy would remain the same: claim anchor first, evidence summary second, stance and confidence near the top, and source inspection always available. This modularity is why the card framework is appropriate as a design contribution rather than a one-off interface mockup.

Limitations

This study has four limitations. First, the UI evaluation uses deterministic proxy metrics rather than a human-subject experiment. The metrics are useful for early design comparison, but they do not replace eye tracking, task-completion studies, or qualitative interviews. The scan-burden and identification-readiness values should therefore be interpreted as interface estimates, not as observed decision time or observed user accuracy.

Second, the evidence-card generator is constrained and extractive. It is LLM-style in structure and output format, but it does not evaluate a live proprietary or open-source generative model. This choice improves stability and source accountability, while future work should compare multiple production LLMs under the same UI framework.

Third, BEIR SciFact qrels evaluate article relevance, not full stance. The support/refute/insufficient badges in the BEIR experiment are generated UI labels. The original SciFact dev validation adds gold stance and rationale checks, but the stance results show that this prototype should not be interpreted as a verified claim-validation system. Fourth, the primary experiment is limited to scientific abstracts. The SciFact-Open candidate-pool stress test broadens the setting, but larger biomedical or multidisciplinary corpora may require different retrieval models, evidence summarization methods, citation grouping, and uncertainty visualization strategies.

A further limitation is that the current confidence cue is calibrated for within-system comparison rather than for probabilistic truth. Users should not read the confidence bar as a probability that the scientific claim is true. It is a visual cue derived from retrieval and evidence-overlap signals. A deployed system would need stronger calibration, explicit user education, and domain-specific thresholds, especially in high-stakes biomedical or policy contexts.

CONCLUSION

This paper presented a UI/UX design framework for LLM-style evidence cards in scientific search interfaces and evaluated it on SciFact and SciFact-Open data. BM25-based baselines remained strong, while the proposed evidence-card pipeline maintained comparable retrieval quality and improved deterministic interface-level evidence visibility and estimated scan burden. The findings show that evidence cards are most valuable when treated as a visual evidence hierarchy: they reveal claim anchoring, evidence summary, stance cue, confidence, ranking reason, and source access in a single inspectable unit. For graphic design and visual communication research, the contribution is a concrete design framework and empirical UI prototype rather than a claim of model supremacy or verified scientific fact checking. The study demonstrates how LLM-style evidence organization can create source-bound explanation components without replacing source inspection. Scientific search interfaces benefit when ranked results are transformed into transparent evidence cards that make support, insufficiency, and uncertainty visible.

REFERENCES

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science, 5, 133-143.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). Readings in information visualization: Using vision to think. Morgan Kaufmann.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2270-2282.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171-4186.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv.

Hearst, M. A. (2009). Search user interfaces. Cambridge University Press.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422-446.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769-6781.

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When(ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 5092-5103.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science, 42(5), 361-371.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuettler, H., Lewis, M., Yih, W.-t., Rocktaeschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-15.

Marchionini, G. (2006). Exploratory search: From finding to understanding. Communications of the ACM, 49(4), 41-46.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81-97.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2014-2037.

Nielsen, J. (1994). Usability engineering. Morgan Kaufmann.

Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. arXiv.

Norman, D. A. (2013). The design of everyday things: Revised and expanded edition. Basic Books.

Pirolli, P., & Card, S. (1999). Information foraging. Psychological Review, 106(4), 643-675.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982-3992.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. Journal of the American Society for Information Science and Technology, 53(2), 145-161.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4), 333-389.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the IEEE Symposium on Visual Languages, 336-343.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285.

Thakur, N., Reimers, N., Rueckle, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 1-17.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 7534-7550.

Wadden, D., Lo, K., Wang, L. L., Cohan, A., & Hajishirzi, H. (2022). SciFact-Open: Towards open-domain scientific claim verification. Findings of the Association for Computational Linguistics: EMNLP 2022, 4719-4734.

Ware, C. (2012). Information visualization: Perception for design (3rd ed.). Morgan Kaufmann.