International Journal of Graphic Design (IJGD)
Volume 4 No. 1 April 2026
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 224-245
DOI: https://doi.org/10.51903/ijgd.v4i1.4002
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Grounded AI Pointer Cards for Browser Agents: Evidence-Aware Web Element Reranking and Action Explanation on WebLINX

Yinchen Shi*[1], Yuxuan Ren[2]
[1] Computer Science, New York University, NY, USA
[2] Chemical Engineering & Data Science, University of Washington, WA, USA
Email Address: ogyinchen@gmail.com

Abstract. *AI browser agents increasingly promise to turn the mouse pointer into a situated interface for selecting page content, asking follow-up questions, scrolling to relevant sections, and opening contextual panels. For graphic design research, this shift matters because the pointer card is both a computational output and a visual communication device: it must make the target, supporting evidence, and next action legible at the moment of use. This paper evaluates evidence-aware web element reranking as the grounding layer for AI pointer cards. The empirical study uses the WebLINX candidate-reranking data, including validation, same-distribution test, and unseen-website test partitions. Each query is paired with candidate DOM/page elements and binary relevance labels. Nine reranking methods were evaluated, ranging from random and DOM-salience controls to BM25 variants, a transparent hybrid formula, a logistic reranker, and a depth-limited decision-tree reranker. Across 5,883 queries and 1,556,655 query–element pairs, the tree reranker achieved the highest same-distribution nDCG@10 (0.272) and Recall@10 (0.473), while the fixed hybrid achieved the highest unseen-website Top-1 accuracy (0.104). Combining lexical and DOM evidence improved ranking, whereas layout features transferred less consistently across websites. Because the first-ranked target was correct in only 10.4% of unseen-website queries, the findings support a confirm-before-act grounding prototype with visible alternatives and correction controls rather than autonomous execution. The reported metrics evaluate target ranking rather than card usability.*

Keywords: *browser agents; web element grounding; pointer cards; learning to rank; human–AI interaction*

INTRODUCTION

The web browser is becoming a site where interface design, information retrieval, and agentic AI meet. Experimental AI pointer systems show this convergence clearly. Google Labs (2026) describes Disco as an environment where a user can select page content, receive specific answers without leaving the webpage, continue the conversation from pointer to side panel, and ask the pointer to scroll, highlight information, and open relevant tabs. Baranes and Marchant (2026) similarly frame the AI-enabled pointer as a prototype that carries context across applications while maintaining the user's flow. These public prototypes make an old design object—the cursor—into a situated conversational interface. The design question is therefore not only whether an agent can propose an action, but whether the interface can show why a particular element on the page was selected.

This paper studies that question through the idea of a grounded AI pointer card. A pointer card is a compact visual object that appears beside, below, or after a pointer interaction. It can propose a click, input, scroll, highlight, or answer while exposing the evidence behind that proposal. For a graphic design journal, the card is a relevant object because it compresses several representational layers: the user's intent, the page's typographic and spatial hierarchy, the ranked

Received: Januari 2026; Revised: February 2026; Accepted: March 2026; Published: April 2026
*Corresponding author, ogyinchen@gmail.com

candidate element, and the explanation that lets the user inspect the proposed action. In this study, the card is a visual translation of ranking evidence rather than a validated usability intervention.

Web element grounding is the technical task beneath this design object. Given a user's instruction and the current browser context, the model must identify which candidate DOM or page element should be selected, highlighted, or used as evidence. The original WebLINX benchmark introduces conversational web navigation with 100,000 interactions and 2,300 expert demonstrations over more than 150 real websites (Lù et al., 2024). The candidate-reranking form of WebLINX turns the grounding problem into a retrieval task: at each step, it ranks relevant elements among many candidates. This setting supplies a proposed target and ordered alternatives for a pointer-card prototype. If the first target is wrong, even a plausible explanation can direct attention to the wrong element. Evidence-grounded interfaces in other high-consequence domains similarly make provenance and verification visible to users (Chen & Xu, 2026; Wu et al., 2023; K. Zhang et al., 2023; Zhong, Wu, et al., 2025). For pointer cards, the analogous requirement is to show why a page element was proposed and to require confirmation before action.

The paper connects each grounding metric to a distinct interface decision. Top-1 accuracy measures whether the primary highlighted element is correct and therefore indicates whether a single-card state is defensible. Recall@5 estimates whether a short alternative stack contains a relevant target, while Recall@10 measures coverage in an expanded candidate pool rather than the usability of displaying ten cards. MRR@10 and nDCG@10 describe how early relevant elements appear when alternatives are inspected. These measures are standard in information retrieval and learning-to-rank research (Järvelin & Kekäläinen, 2002; T.-Y. Liu, 2009; Y. Wang et al., 2013), but they do not measure explanation readability, confirmation behavior, or correction success. A low Top-1 score therefore favors confirmation before action; shortlist recall supports optional alternatives; and weak list ordering favors cycling or manual target selection rather than autonomous execution.

The paper asks three research questions. First, which lightweight evidence sources—visible text, DOM attributes, structural salience, and layout cues—are most useful for identifying the correct page element? Second, how do transparent lexical and hybrid methods compare with supervised rerankers on same-distribution and unseen-website splits? Third, what design implications for evidence presentation, confirmation, and correction can be drawn from the observed rankings? The answers are based on query-grouped metrics from validation, test_iid, and test_web. Figure 1 summarizes the study's flow from user intent and browser context to ranked evidence and prototype presentation.

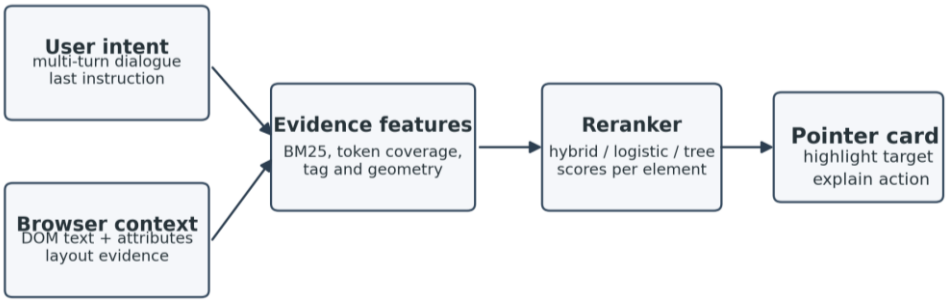

Figure 1. Pipeline connecting user intent, browser context, evidence features, reranking, and pointer-card presentation.

LITERATURE REVIEW

Browser agents have developed from question-answering systems that use a browser as a tool into agents that must perceive, ground, and act within dynamic interfaces. WebGPT demonstrated that browser-assisted question answering could be improved through feedback and browsing traces, showing the value of connecting language models to web evidence rather than relying only on parametric memory (Nakano et al., 2021). Mind2Web extended this idea to generalist web tasks, collecting open-ended instructions across many websites and domains and revealing the difficulty of transferring action policies across real interfaces (Deng et al., 2023). WebArena then emphasized realistic web environments with executable tasks, while VisualWebArena added multimodal visual grounding for tasks in which screenshots and spatial layout are essential (Koh et al., 2024; S. Zhou et al., 2024). SeeAct likewise showed that even strong multimodal models require careful grounding when acting on webpages (B. Zheng et al., 2024). These studies frame browser agency as a grounding problem: the model must bind language to an interface element before the interface can propose an element-specific action.

WebLINX is particularly relevant because it treats conversational web navigation as multi-turn interaction rather than a single isolated click. Lù et al. (2024) argue that the volume of page information makes it impractical for many models to process full webpages in real time, so relevant element ranking becomes a necessary pruning stage. The candidate-reranking task used here follows the same logic. It supplies the proposed target and ordered alternatives for the pointer-card prototype. Because the model scores were not calibrated as probabilities (Xin, 2025b), they are used for ordering rather than as claims of confidence.

Related trustworthy-agent work makes the same intermediate layer explicit: retrieval must be paired with evidence-chain scoring, source attribution, and coverage-aware abstention when the system cannot support a claim (Jin, 2025a; C. Li et al., 2024; Su, Chen, et al., 2026; Su et al., 2025; Zhong, Chen, et al., 2025). For a browser pointer card, the analogous failure is not only a wrong answer but a wrong highlighted element with a plausible-looking explanation.

Information retrieval supplies methods and metrics for this layer. BM25 remains a strong and interpretable lexical baseline because it combines term frequency, document length normalization, and inverse document frequency in a transparent relevance score (Robertson & Zaragoza, 2009). BEIR and MTEB have shown that retrieval and embedding models require diverse benchmarks because performance varies by task, domain, and evaluation setting (Muennighoff et al., 2023; Thakur et al., 2021; R. Zhang et al., 2025). In the present setting, the elements are not full documents; they are small fragments of visible text, attributes, tags, and bounding boxes. That makes retrieval evidence more fragmented and makes fielding important. An input element can be relevant because of an attribute such as a field hint or ID even when it has little visible text. A heading can be relevant through inner text. A button can be relevant through its role, tag, or spatial context.

The design literature helps explain why a pointer card must expose this evidence carefully. Bertin (1983) described visual variables as a language for encoding relations, while Tufte (2001) argued that visual displays should maximize meaningful evidence and minimize chartjunk. Ware (2013) connects such choices to perception, showing that position, contrast, and grouping guide attention before conscious reading. Card et al. (1999) and Heer and Shneiderman (2012) further position information visualization as a way to support reasoning through interaction. A pointer card uses the same principles at a micro scale: the target outline, the evidence label, and the action verb should communicate a coherent relation between the user's request and the page element.

Recent card-based interface studies make this micro-scale problem more concrete. Evidence cards for scientific search, disclosure review cards for financial reports, risk and data-integrity cards for LLM agents, incident visualization cards for cloud logs, and trust-calibrated product-recommendation cards all treat the card as a compact evidence hierarchy rather than as decorative output (Jin, 2025b; Nie et al., 2026; Su, Rao, et al., 2026; B. Zhang et al., 2025; Zhong, Wu, et al., 2025). Design-rationale, visual-brief, and dashboard frameworks likewise show how short textual rationales can be made editable, inspectable, and aligned with visual hierarchy (Y. Li et al., 2025; Z. Li et al., 2025; G. Liu et al., 2025; Mu et al., 2026; Tu et al., 2025; Wu et al., 2025).

Human-computer interaction research adds constraints on trust and usability. Norman (2013) emphasizes affordances, feedback, and mappings; Nielsen (1994) emphasizes visibility of system status and error prevention. Human-AI interaction guidelines recommend that AI systems make clear what they can do, show contextually relevant information, and support efficient correction when the system is wrong (Amershi et al., 2019; Sultana et al., 2025). Shneiderman (2022) argues for human-centered AI that keeps people in control through understandable interaction. These principles are consistent with explainable AI research, where local explanations

are useful when they help users evaluate a specific prediction or action rather than merely decorate a model output (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016). A pointer card should therefore be designed as actionable evidence, not as a generic confidence badge.

Safety-oriented LLM studies extend this point by showing that self-checking, guardrail updates, selective refusal, and adversarial verification are interaction problems as much as modeling problems (C. Li et al., 2023; Nie & Zheng, 2023; D. Zheng & Li, 2024; D. Zheng et al., 2024; B. Zhou, Jin, et al., 2025). In the pointer-card setting, these ideas motivate cards that can abstain, ask for confirmation, or show alternatives when grounding evidence is weak. The present study joins these strands by evaluating the grounding layer and translating the results into card-level design requirements. The empirical comparison uses lightweight, inspectable methods so that each ranking can be traced to lexical, DOM, or layout evidence. The methods are implemented with standard machine-learning tools (Pedregosa et al., 2011), allowing the stability of each evidence type to be compared across unfamiliar websites.

These design requirements also connect to explainable recommendation and reranking work, where a ranked item should be accompanied by faithful evidence rather than post hoc persuasion. Personalized reranking, review-grounded recommendation, and off-policy evaluation of slot policies are useful analogues for pointer-card stacks because each asks whether a top candidate, a shortlist, and its explanation can remain trustworthy under changing user context (Chang et al., 2026; Chang et al., 2023; Meng et al., 2026; Mu et al., 2025; Ye et al., 2026; Y. Zhang & Zhang, 2025a).

METHODS

The empirical evaluation used three WebLINX candidate reranking partitions: validation, test_iid, and test_web. The validation partition was used for model fitting and parameter choice; test_iid measured same-distribution generalization; test_web measured performance on unseen websites. Each partition contained a query file, a candidate-element corpus file, a relevance-label file, and a top-ranked candidate list. The files were joined by query ID and corpus ID so that each evaluated row represented one query-element pair with a binary relevance label. The joins preserved query-level grouping, which is essential because all metrics were computed over ranked candidate lists rather than over independent binary classifications.

Table 1. Dataset inventory by split.

Split	Queries	Candidate elements	Positive labels	Mean candidates/query	Median candidates/query
Validation	1,301	316,508	1,315	243.3	199.0
Test IID	1,438	405,972	1,514	282.3	211.5
Test Web	3,144	834,175	3,329	265.3	238.0

Table 1 reports the scale of the evaluation. The study covers 5,883 queries, 1,556,655 labeled query-element pairs, and 6,158 positive labels. The average query had between 243 and 282 candidate elements, which confirms that pointer grounding is not a trivial classification problem. Even a high-quality card must be produced after ranking hundreds of possible page elements at a single interaction step. Figure 2 summarizes the same scale visually.

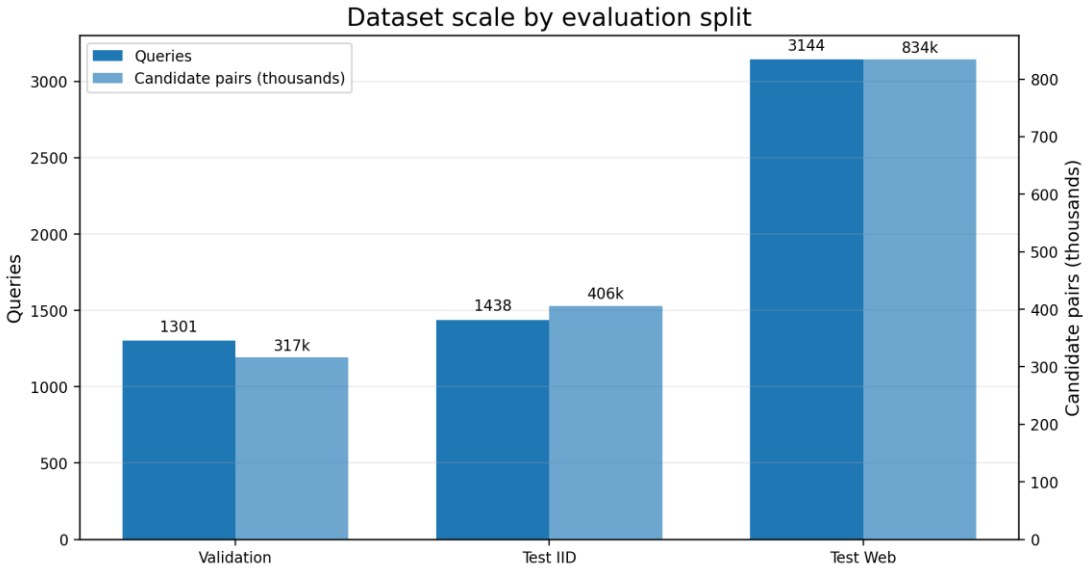

Figure 2. Scale of queries and candidate query-element pairs in the evaluation data.

The element text field was parsed into tag, XPath, bounding-box coordinates, attributes, children text, and inner text. The parser extracted numeric geometry when available and treated missing text fields as empty strings rather than as model evidence. Text was normalized with lowercase tokenization and alphanumeric filtering. The resulting features included query length, last-user-turn length, turn count, character lengths for inner text, attributes and children, XPath depth, x and y position, width, height, log area, binary flags for visible text and actionability, tag indicators for common elements, and retrieval-style text scores. Table 2 shows that only 14.5% to 17.4% of candidate elements contained visible text, while about 18.3% to 19.6% were actionable. This sparsity supports the use of DOM attributes and layout cues in addition to visible labels.

Table 2. Element-field and layout statistics.

Split	Mean element tokens	Median tokens	Actionable share	Visible text share	Mean bbox area	Median XPath depth
Validation	17.52	15.0	0.191	0.174	207,022	15.0
Test IID	17.87	15.0	0.196	0.145	174,830	16.0
Test Web	17.40	14.0	0.183	0.162	3,865,578	17.0

Tag distribution also matters for design interpretation because candidate sets are dominated by layout containers rather than obvious buttons or links. Table 3 shows that div and span

elements account for large shares of candidates across all splits, while anchors, table cells, rows, images, inputs, and buttons appear with smaller shares. A pointer-card system must therefore avoid treating all DOM nodes as equally communicative. The interface should favor elements whose evidence can be explained: a visible label, an action role, a field id, or a bounding box that matches the user's visual attention.

Table 3. Dominant candidate tags by count and share.

Tag	Validation	Test IID	Test Web
div	138,826 (0.439)	178,538 (0.440)	389,513 (0.467)
span	42,208 (0.133)	54,533 (0.134)	127,234 (0.153)
a	23,504 (0.074)	23,103 (0.057)	53,841 (0.065)
path	20,337 (0.064)	27,385 (0.067)	44,801 (0.054)
svg	17,316 (0.055)	20,933 (0.052)	41,570 (0.050)
li	14,850 (0.047)	16,080 (0.040)	26,531 (0.032)
button	12,343 (0.039)	16,981 (0.042)	32,281 (0.039)
img	7,176 (0.023)	6,822 (0.017)	13,987 (0.017)
td	0 (0.000)	14,543 (0.036)	12,796 (0.015)

Nine reranking methods were compared, as summarized in Table 4. The random control used deterministic pseudorandom scores to verify that the metrics behaved as expected. DOM salience scored actionability, visible text, tag type, and geometry without using query tokens. Token overlap compared the last user instruction with element text and attributes. Three BM25 variants were evaluated: full-dialogue BM25, last-turn BM25, and fielded BM25 over inner text, attributes, and children. The fixed hybrid combined last-turn BM25, token coverage, actionability, visible text, and spatial salience into a transparent score. Two supervised rerankers were trained on validation pairs: an elastic-net logistic classifier using stochastic gradient descent and a depth-limited decision tree. The tree was constrained to a maximum depth of 9 and minimum leaf size of 20 to keep its behavior inspectable.

Table 4. Reranking methods and evidence sources.

Method	Evidence used	Role in comparison
Random control	Deterministic pseudorandom scores	Sanity check for group-wise ranking metrics
DOM salience	Actionability, tag type, text presence, geometry	Tests whether visual and structural cues alone localize useful elements
Token overlap	Last-user tokens matched against element text and attributes	Simple lexical pointer grounding
BM25 dialogue	BM25 over the full dialogue context	Information-retrieval baseline for conversational context
BM25 last turn	BM25 using the most recent user instruction	Tests short-context intent matching
Fielded BM25	Weighted BM25 over inner text, attributes, and children	Separates visible label text from hidden DOM descriptors
Fixed hybrid	Weighted blend of BM25, overlap, and DOM salience	A transparent ranking formula designed for pointer cards
Logistic reranker	Elastic-net SGD classifier trained on validation pairs	Lightweight supervised score model
Tree reranker	Depth-limited decision tree trained on validation pairs	Nonlinear evidence interaction model with readable feature importance

Auditability is also important in operational retrieval systems, where operators need to inspect the evidence behind document and incident recommendations (C. Li et al., 2026; G. Liu et al., 2024; Xin, 2025a; B. Zhang et al., 2024). The present comparison therefore favors features whose contributions can be traced to text, attributes, or geometry. Evaluation was grouped by query. For each query, candidates were sorted by model score, with a deterministic hash used only to break exact ties. Metrics included Top-1 accuracy, MRR@10, nDCG@10, Recall@5, and Recall@10. Top-1 evaluates the proposed primary target; Recall@5 evaluates coverage in a short alternative stack; Recall@10 evaluates coverage in an expanded candidate pool; and MRR@10 and nDCG@10 evaluate the ordering of those alternatives. These metrics characterize grounding performance but do not measure card readability or correction usability. Runtime logging recorded feature construction, model training, and prediction time. Figure 3 summarizes nDCG@10 across the nine methods and three evaluation splits; Tables 5–7 report the complete split-specific metrics.

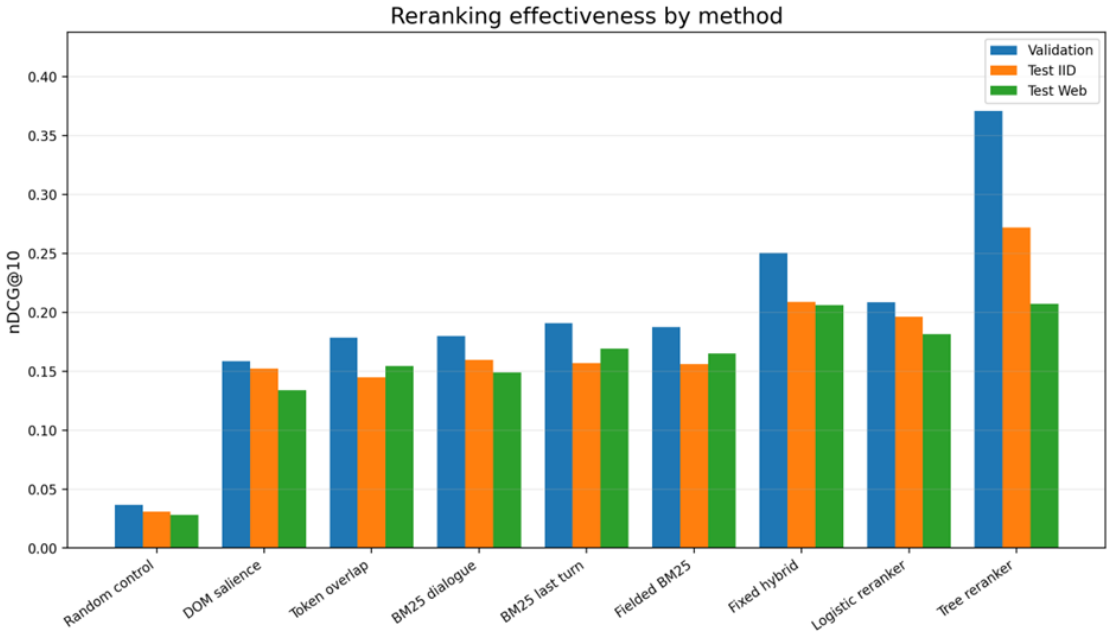

Figure 3. nDCG@10 by reranking method and evaluation split.

RESULTS

Table 5 summarizes performance on the model-fitting split. Because the supervised rerankers were fitted and tuned on validation pairs, their validation values describe fitted-split behavior rather than independent generalization. The random control reached 0.036 nDCG@10, DOM salience reached 0.158, and BM25 last turn reached 0.191 nDCG@10 and 0.108 Top-1. The fixed hybrid reached 0.250 nDCG@10. The tree reranker produced the highest validation

values, with 0.371 nDCG@10, 0.160 Top-1, and 0.639 Recall@10. Same-distribution testing, shown in Table 6, preserves the broad ordering while lowering absolute scores.

Table 5. Validation reranking results.

Method	Top-1	MRR@10	nDCG@10	Recall@5	Recall@10
Random control	0.010	0.024	0.036	0.042	0.079
DOM salience	0.068	0.118	0.158	0.168	0.294
Token overlap	0.098	0.150	0.179	0.219	0.273
BM25 dialogue	0.093	0.144	0.180	0.209	0.299
BM25 last turn	0.108	0.162	0.191	0.241	0.284
Fielded BM25	0.105	0.158	0.188	0.229	0.283
Fixed hybrid	0.141	0.206	0.250	0.289	0.396
Logistic reranker	0.075	0.156	0.208	0.263	0.380
Tree reranker	0.160	0.289	0.371	0.472	0.639

The tree reranker remained the strongest method on test_iid with 0.108 Top-1, 0.211 MRR@10, 0.272 nDCG@10, 0.353 Recall@5, and 0.473 Recall@10. The correct target appeared in the top 10 for 47.3% of queries, but the first target was correct for only 10.8%. This pattern supports an alternative-recovery mechanism rather than unconfirmed execution. The fixed hybrid reached 0.209 nDCG@10 and 0.104 Top-1. Logistic reranking improved recall relative to the BM25 variants but did not improve Top-1, suggesting that its ranking score was useful for forming candidate pools but less decisive for selecting a primary target.

Table 6. Same-distribution test_iid reranking results.

Method	Top-1	MRR@10	nDCG@10	Recall@5	Recall@10
Random control	0.003	0.019	0.031	0.035	0.071
DOM salience	0.054	0.108	0.152	0.166	0.300
Token overlap	0.076	0.119	0.145	0.181	0.231
BM25 dialogue	0.080	0.128	0.160	0.190	0.272
BM25 last turn	0.090	0.133	0.157	0.192	0.235
Fielded BM25	0.084	0.130	0.156	0.195	0.242
Fixed hybrid	0.104	0.167	0.209	0.241	0.348
Logistic reranker	0.084	0.155	0.196	0.252	0.329
Tree reranker	0.108	0.211	0.272	0.353	0.473

The unseen-website results in Table 7 provide the strongest test of transfer to unfamiliar sites. Test_web contains 3,144 queries and 834,175 candidate pairs, making it the largest and hardest split. The fixed hybrid achieved the highest Top-1 accuracy at 0.104, while the tree reranker achieved the highest nDCG@10 at 0.207 and Recall@10 at 0.346. BM25 last turn reached 0.099 Top-1, indicating that recent user intent can be more useful than the whole dialogue when choosing a visible target. The logistic reranker reached 0.343 Recall@10 but only 0.058 Top-1. Even the highest unseen-website Top-1 value means that 89.6% of primary highlights were incorrect, and the best Recall@10 value shows that the expanded candidate pool remained incomplete. No evaluated method therefore supports autonomous execution: the first target should

be presented as a suggestion requiring confirmation, with alternatives and manual correction available. Figure 4 compares the strongest methods on Top-1 and Recall@10.

Table 7. Unseen-website test_web reranking results.

Method	Top-1	MRR@10	nDCG@10	Recall@5	Recall@10
Random control	0.007	0.019	0.028	0.031	0.062
DOM salience	0.030	0.090	0.134	0.162	0.281
Token overlap	0.080	0.128	0.154	0.189	0.244
BM25 dialogue	0.065	0.115	0.149	0.167	0.272
BM25 last turn	0.099	0.144	0.169	0.202	0.255
Fielded BM25	0.091	0.139	0.165	0.203	0.251
Fixed hybrid	0.104	0.166	0.206	0.246	0.338
Logistic reranker	0.058	0.132	0.181	0.229	0.343
Tree reranker	0.094	0.165	0.207	0.258	0.346

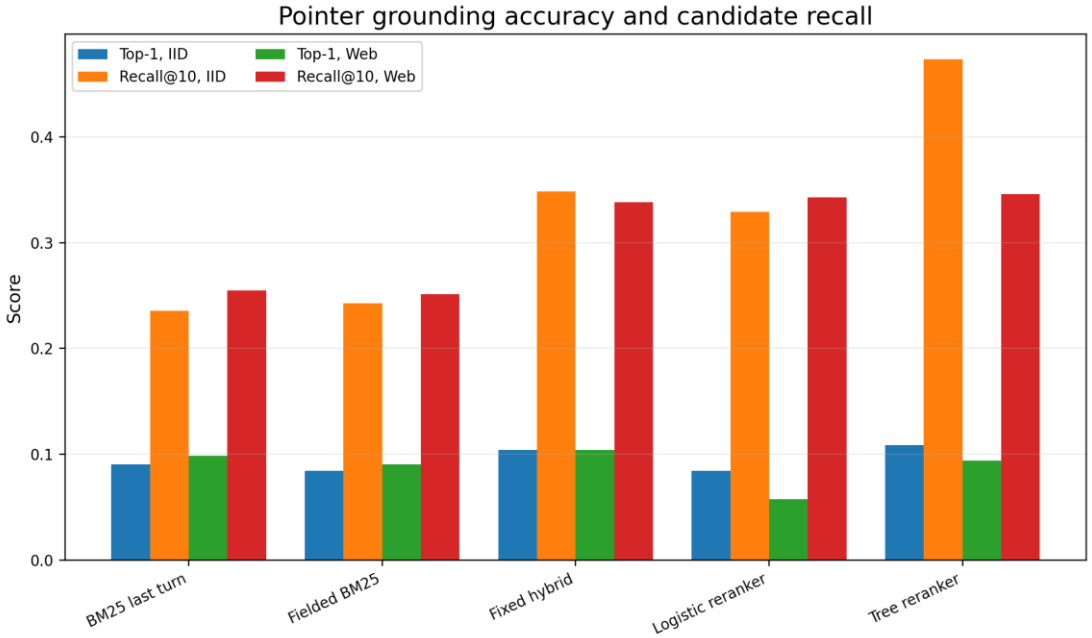

Figure 4. Top-1 grounding accuracy and Recall@10 for the strongest test methods.

Table 8. Cross-website nDCG@10 generalization gap.

Method	Validation nDCG@10	Test IID nDCG@10	Test Web nDCG@10	IID-Web gap	Relative gap
BM25 dialogue	0.180	0.160	0.149	0.011	6.8%
BM25 last turn	0.191	0.157	0.169	-0.013	-8.0%
DOM salience	0.158	0.152	0.134	0.018	12.1%
Fielded BM25	0.188	0.156	0.165	-0.009	-5.7%
Fixed hybrid	0.250	0.209	0.206	0.003	1.5%
Logistic reranker	0.208	0.196	0.181	0.015	7.5%
Random control	0.036	0.031	0.028	0.003	8.6%
Token overlap	0.179	0.145	0.154	-0.009	-6.5%
Tree reranker	0.371	0.272	0.207	0.065	23.9%

Table 8 and Figure 5 isolate cross-website robustness. The tree reranker had the largest positive nDCG gap, dropping from 0.272 on test_iid to 0.207 on test_web, a 23.9% relative

decrease. The tree retained the highest unseen-website nDCG@10, but the decrease indicates lower transfer stability when website conventions change. By comparison, the fixed hybrid's nDCG@10 gap was 0.003, or 1.5%. Fielded BM25, token overlap, and BM25 last turn were slightly stronger on the unseen-website split than on test_iid, suggesting that lexical intent matching can generalize when element labels and attributes are direct.

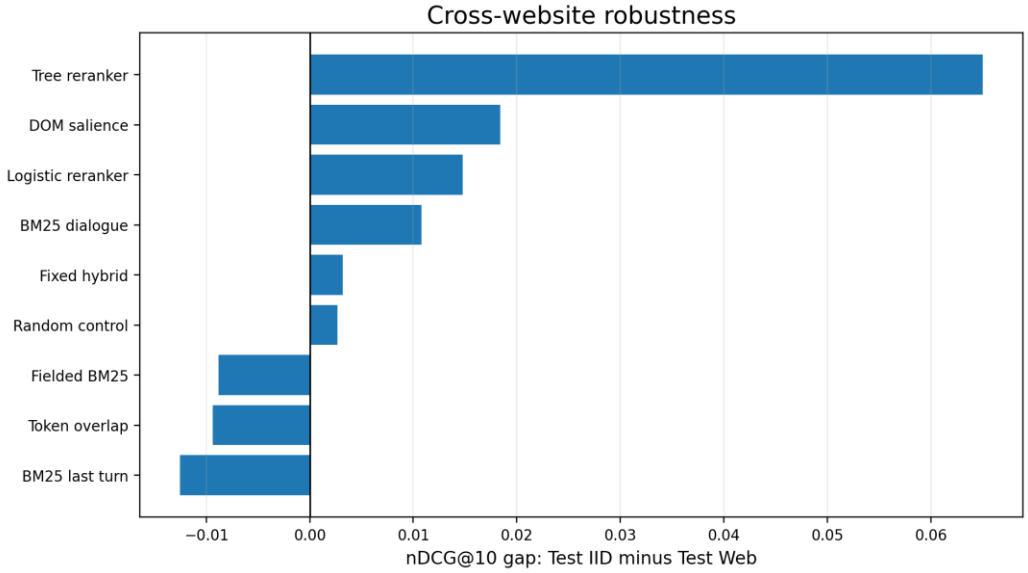

Figure 5. nDCG@10 generalization gap from same-distribution to unseen-website testing.

The ablation study in Table 9 and Figure 6 explains why the best scores require mixed evidence. On test_iid, text-only features reached 0.177 nDCG@10, DOM and layout only reached 0.227, and text plus DOM without layout reached 0.270. The full feature set produced similar nDCG@10 with higher Recall@10. On test_web, text plus DOM without layout reached the strongest ablation nDCG@10 at 0.242, while adding all layout features reduced nDCG@10 to 0.212 under the ablation tree protocol. Geometry improved same-distribution candidate recall but transferred inconsistently across websites when absolute coordinates were used. Spatial evidence remains necessary for placing a highlight, but textual or semantic evidence should justify why that element was proposed.

Table 9. Ablation results for evidence families.

Split	Ablation	Features	Top-1	nDCG@10	Recall@10
Test IID	Text scores only	9	0.095	0.177	0.281
Test IID	DOM and layout only	30	0.086	0.227	0.405
Test IID	Text + DOM, no layout	35	0.134	0.270	0.445
Test IID	All features	40	0.102	0.269	0.475
Test Web	Text scores only	9	0.087	0.171	0.271
Test Web	DOM and layout only	30	0.078	0.172	0.280
Test Web	Text + DOM, no layout	35	0.117	0.242	0.385
Test Web	All features	40	0.102	0.212	0.343

The feature-importance results in Table 10 and Figure 7 support the same interpretation. The tree assigned the largest importance to the fixed hybrid score (0.358), followed by width (0.158), log area (0.081), full-dialogue BM25 (0.065), y position (0.059), x position (0.048), inner-character length (0.045), and XPath depth (0.042). These values describe what the model used, not which fields should be shown to users. A pointer card can expose the interpretable subset—element type, matched label or attribute, and approximate location—while omitting raw feature values.

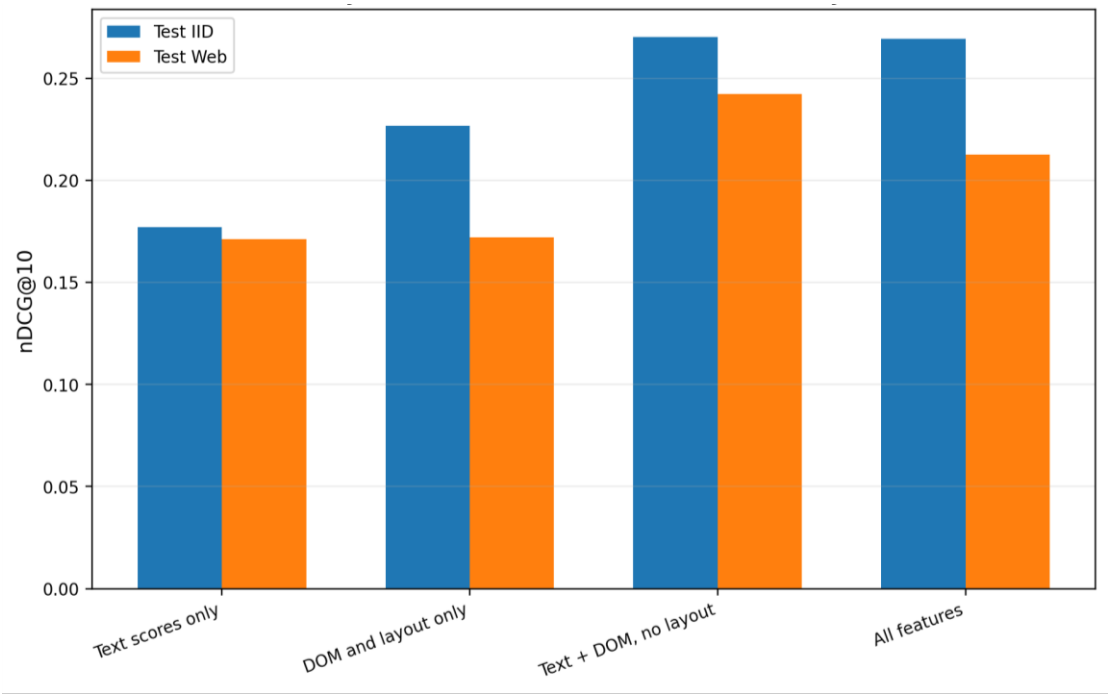

Figure 6. Ablation comparison for text, DOM, and layout evidence.

Table 10. Leading feature importances for the tree reranker.

Feature	Importance
fixed_hybrid	0.358
width	0.158
log_area	0.081
bm25_full_dialogue	0.065
y	0.059
x	0.048
inner_char_len	0.045
xpath_depth	0.042

Table 11 reports the logged batch runtime for feature construction, fitting, and prediction. Feature construction took 21.2 seconds for validation, 26.0 seconds for test_iid, and 51.4 seconds for test_web, or approximately 16–18 milliseconds per query group in the logged run. Logistic training took 5.015 seconds and tree training took 2.242 seconds. On test_web, prediction after

feature construction took 0.324 seconds for the logistic reranker and 0.029 seconds for the tree reranker over 834,175 candidate pairs.

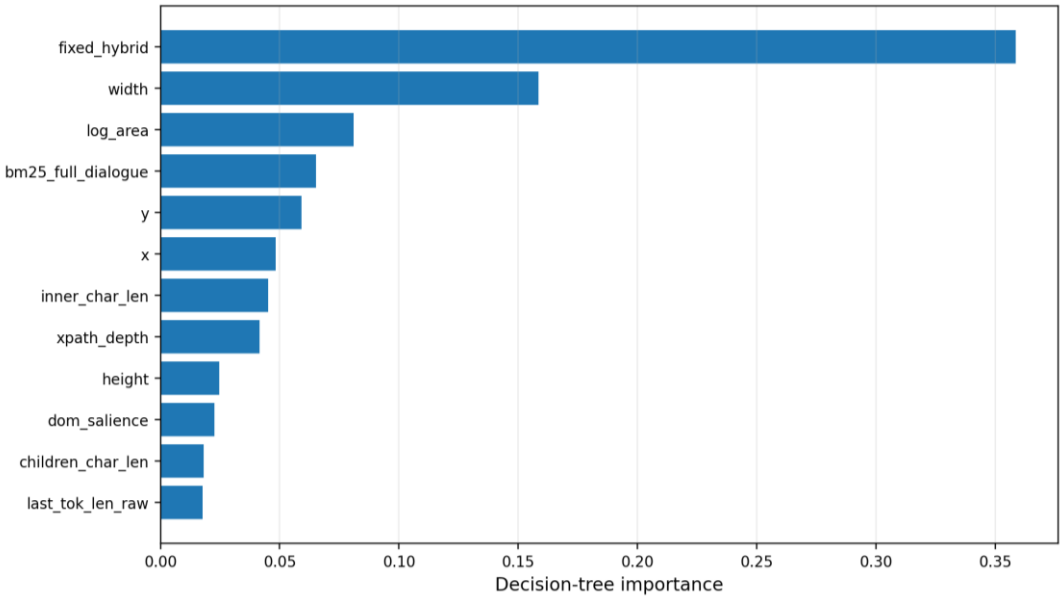

Figure 7. Most influential features in the decision-tree reranker.

These measurements support offline analysis and prototype scoring, but they are not end-to-end browser-latency measurements. Rendering order and structural consistency remain separate front-end evaluation concerns. Work on layout-aware progressive PDF rendering and vision-language web prototype evaluation shows that perceived readiness depends on which components or structural cues become usable first (Chen & Li, 2025; H. Wang et al., 2025; B. Zhou, Wang, et al., 2025).

Table 11. Runtime and deployment profile.

Component	Seconds	ms/query
Feature construction: validation	21.200	16.295
Feature construction: test_iid	26.000	18.081
Feature construction: test_web	51.400	16.349
Logistic reranker training	5.015	
Tree reranker training	2.242	
Logistic reranker prediction: test_web	0.324	0.103
Tree reranker prediction: test_web	0.029	0.009

Table 12 presents four successful unseen-website cases and shows how ranked evidence can be expressed as compact pointer-card text. Each card uses an action verb, target element type, visible or DOM evidence, and a concise rationale (Y. Zhang & Zhang, 2025b). The values in the final column are uncalibrated ranking scores; they determine candidate order but do not represent probabilities of correctness.

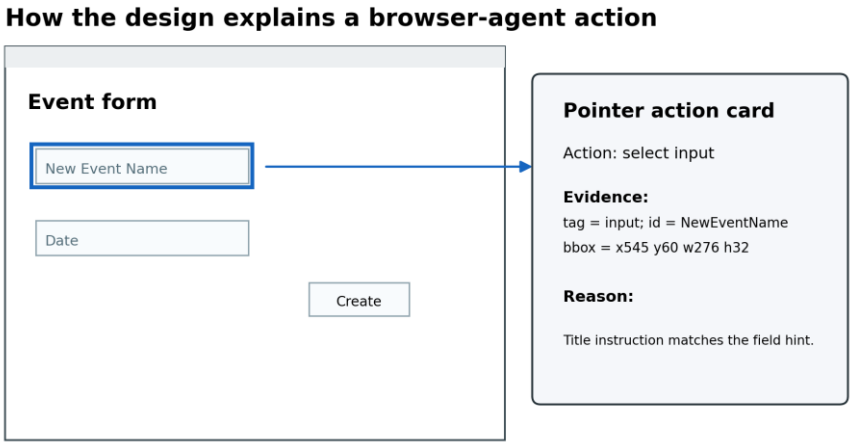

Figure 8. Pointer-card composition linking a highlighted element to concise evidence.

Figure 8 shows the prototype composition linking a proposed highlight to a small evidence block. This composition is intended to support inspection and correction, but the reranking results do not establish its readability or usability. This distinction follows human-AI interaction guidance to show status, support correction, and keep the person in control (Amershi et al., 2019; Shneiderman, 2022). Because pointer-card explanations are brief, their microcopy should state the proposed action plainly and leave uncertainty open to correction, consistent with interface-text research on manipulative wording (Xu et al., 2025).

Table 12. Pointer-card wording for four successful unseen-website cases.

User intent	Predicted target evidence	Action-card text	Ranking score
Let us use "Team Building Workshop" as the title.	input; id=NewEventName; value=New Event Name; bbox x=545 y=60 w=276 h=32	Select the title input because its id and value match the workshop title instruction.	0.971
Please provide the names of the newest 3 books.	h3; text=13 New Books Coming in May; bbox x=274 y=259 w=470 h=25	Highlight the heading because it names the newest-book section.	0.971
Open the "Food" category and show me some options.	a; text=Food; bbox x=716 y=112 w=33 h=24	Select the Food link because its visible label matches the category request.	0.947
Scroll down and locate the "Follow" button.	button; text=Follow; bbox x=42 y=430 w=68 h=22	Select the Follow button because the visible label matches the locate request.	0.947

DISCUSSION

The ranking results support a confirmation-first interface rather than an autonomous pointer. On unseen websites, the best Top-1 accuracy was 0.104 and the best Recall@10 was 0.346; neither a primary target nor a ten-item candidate pool was reliable enough to execute without review. Table 13 translates each ranking signal into a bounded interface behavior. A primary card should ask the user to confirm the highlighted target, a compact alternative stack

should remain available, and the user should be able to cycle through, select, or reject candidate highlights before an action occurs.

Table 13. Mapping from grounding evidence to bounded pointer-card behavior.

Grounding evidence	Pointer-card behavior	Interpretation boundary
Top-1 accuracy	Show the first element as a proposed primary highlight and require confirmation before execution.	Measures first-rank correctness; it does not establish autonomy or card usability.
Recall@5	Offer a compact alternative stack or cycle through up to five candidate highlights.	Measures shortlist coverage; it does not show whether five cards are readable.
Recall@10	Provide an expanded recovery view or manual candidate browser.	Best unseen-web Recall@10 was 0.346, so the relevant target was absent in 65.4% of queries.
MRR@10 and nDCG@10	Order alternatives and place stronger evidence earlier.	Measure ranked position, not probability of correctness.
Raw ranking score	Use to order candidates; display as confidence only after calibration.	The scores in Table 12 are uncalibrated.
Text, DOM, and layout evidence	Pair the highlight with a concise label, attribute, role, or location cue.	Geometry locates a target; text or semantics should explain why it was proposed.

Card-level evaluation requires criteria separate from ranking metrics. Table 14 defines a 0–2 rubric for explanation faithfulness, readability and action clarity, confirmation and uncertainty, and correction support (C. Li et al., 2025). The rubric distinguishes what ranking can supply from what must be judged at the interface level.

Table 14. Card-level evaluation rubric for pointer-card prototypes.

Criterion	0	1	2
Evidence faithfulness	Rationale is unsupported or conflicts with the target.	Evidence is relevant but incomplete or indirect.	Rationale cites visible text, an attribute, or a role tied to the proposed target.
Readability and action clarity	Action or target is unclear or visually dense.	Meaning is understandable but requires inference.	A concise action verb and target are immediately scannable.
Confirmation and uncertainty	Card implies certainty and offers no confirmation.	Either suggestion status or a confirmation control is present.	Card is clearly a proposal with confirm and cancel controls before execution.
Correction support	No recovery path is available.	One alternative or a manual-selection option is available.	User can cycle or select alternatives, choose another element, or decline the action.

No participant-based evaluation was conducted, so the results do not establish comprehension, trust, task speed, visual preference, or correction success. Table 12 and Figure 8

specify how ranking evidence can be shown, while Table 14 defines what a card-level assessment should measure. The study is also limited to WebLINX candidate reranking and lightweight text, DOM, and layout features; raw scores were not calibrated as correctness probabilities.

CONCLUSION

This paper evaluated evidence-aware web element reranking as the grounding layer for AI pointer cards. Across WebLINX, the tree reranker produced the strongest same-distribution nDCG@10 and Recall@10, while the fixed hybrid achieved the highest unseen-website Top-1 accuracy and the smallest nDCG@10 distribution gap. Yet the best unseen-website Top-1 accuracy was only 0.104, so the evidence supports a confirmation-first grounding prototype rather than autonomous browser action or validated pointer-card effectiveness. The appropriate interface response is to present a proposed target with traceable text or DOM evidence, keep alternatives accessible, and let the user confirm, replace, or reject the target before execution. Layout features remain useful for locating a highlight, but their cross-site instability makes them unsuitable as the sole justification for an action.

FUTURE RESEARCH

Future work should evaluate the card-level rubric with users or expert interface reviewers, measuring target comprehension, readability, confirmation accuracy, correction success, and task time. Calibration and risk–coverage analysis (Xin, 2026) could then determine when to display a single proposed target, an alternative stack, or an abstention prompt. Multimodal studies that combine WebLINX screenshots with normalized geometry could also test whether perceptual features improve grounding across unfamiliar visual layouts.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

Baranes, A., & Marchant, R. (2026, May 12). Reimagining the mouse pointer for the AI era. Google DeepMind. https://deepmind.google/blog/ai-pointer/

Bertin, J. (1983). Semiology of graphics: Diagrams, networks, maps. University of Wisconsin Press.

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

Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. JEECS

(Journal of Electrical Engineering and Computer Sciences), 11(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

Chang, X., Ye, T., & Luo, S. (2023). LLM-as-reranker for personalized recommendation: Popularity bias mitigation and faithful natural-language explanations on MovieLens 100K. Journal of Advanced Computing Systems, 3(8), 61–78. https://doi.org/10.69987/JACS.2023.30806

Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. Journal of Technology Informatics and Engineering, 4(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490

Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. International Journal of Graphic Design, 4(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., & Su, Y. (2023). Mind2Web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems, 36, 28091–28114.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning [Preprint]. arXiv. https://arxiv.org/abs/1702.08608

Google Labs. (2026). Disco. Retrieved July 23, 2026, from https://labs.google/disco

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. Communications of the ACM, 55(4), 45–54. https://doi.org/10.1145/2133806.2133821

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422–446. https://doi.org/10.1145/582415.582418

Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. Journal of Technology Informatics and Engineering, 4(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. International Journal of Graphic Design, 3(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M. C., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., & Fried, D. (2024). VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 881–905). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.50

Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM

applications. Journal of Advanced Computing Systems, 4(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. Journal of Technology Informatics and Engineering, 5(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

Li, C., Su, W., & Zhang, E. (2023). Lightweight hallucination firewall for enterprise LLM applications: Evidence consistency, self-checking, and small-model detection on TruthfulQA. Journal of Advanced Computing Systems, 3(1), 49–65. https://doi.org/10.69987/JACS.2023.30104

Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. International Journal of Graphic Design, 3(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. International Journal of Graphic Design, 3(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. International Journal of Graphic Design, 3(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. International Journal of Graphic Design, 3(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. Journal of Advanced Computing Systems, 4(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

Liu, T.-Y. (2009). Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3(3), 225–331. https://doi.org/10.1561/1500000016

Lù, X. H., Kasner, Z., & Reddy, S. (2024). WebLINX: Real-world website navigation with multi-turn dialogue. In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 33007–33056). PMLR. https://proceedings.mlr.press/v235/lu24e.html

Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. Journal of Technology Informatics and Engineering, 5(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable

graphic design cards. International Journal of Graphic Design, 4(1), 192–208. https://doi.org/10.51903/ijgd.v4i1.3702

Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small). Journal of Technology Informatics and Engineering, 4(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

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. https://doi.org/10.18653/v1/2023.eacl-main.148

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2021). WebGPT: Browser-assisted question-answering with human feedback [Preprint]. arXiv. https://arxiv.org/abs/2112.09332

Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. International Journal of Graphic Design, 4(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

Nie, J., & Zheng, D. (2023). Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal. Journal of Advanced Computing Systems, 3(1), 66–80. https://doi.org/10.69987/JACS.2023.30105

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

Norman, D. A. (2013). The design of everyday things. Basic Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

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. https://doi.org/10.1145/2939672.2939778

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

Shneiderman, B. (2022). Human-centered AI. Oxford University Press. https://doi.org/10.1093/oso/9780192845290.001.0001

Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. Journal of Technology Informatics and Engineering, 5(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. Journal of Technology Informatics and Engineering, 4(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. International Journal of Graphic Design, 4(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

Sultana, M., Sumaiya, U. K., Saha, R. R., & Tudu, R. (2025). Critical Speculations In AI Design: Ethical Narratives And Visual Experiments In Post-Human Aesthetic Practices. International Journal of Graphic Design, 3(1), 81–101. https://doi.org/10.51903/IJGD.V3I1.2873

Thakur, N., Reimers, N., Rücklé, 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.

Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. International Journal of Graphic Design, 3(1), 210–226. https://doi.org/10.51903/ijgd.v3i1.3714

Tufte, E. R. (2001). The visual display of quantitative information (2nd ed.). Graphics Press.

Wang, H., Ren, Y., & Chang, X. (2025). Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD. Journal of Technology Informatics and Engineering, 4(2), 425–446. https://doi.org/10.51903/jtie.v4i2.523

Wang, Y., Wang, L., Li, Y., He, D., Chen, W., & Liu, T.-Y. (2013). A theoretical analysis of NDCG ranking measures. Proceedings of Machine Learning Research, 30, 25–54. https://proceedings.mlr.press/v30/Wang13.html

Ware, C. (2013). Information visualization: Perception for design (3rd ed.). Morgan Kaufmann.

Wu, Q., Bai, J., & Zhou, X. (2023). Evidence-grounded financial RAG: Reducing numerical hallucination in LLM-generated corporate risk memos. Journal of Advanced Computing Systems, 3(3), 65–84. https://doi.org/10.69987/JACS.2023.30306

Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision

cards. International Journal of Graphic Design, 3(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. Emerging Information Science and Technology, 6(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

Xin, Q. (2025b). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). Journal of Technology Informatics and Engineering, 4(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

Xin, Q. (2026). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. AVITEC, 8(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. Journal of Technology Informatics and Engineering, 4(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). Journal of Technology Informatics and Engineering, 5(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. Journal of Advanced Computing Systems, 4(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. International Journal of Graphic Design, 3(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

Zhang, K., Meng, S., & Zhou, E. (2023). Evidence-grounded trading desk risk memos over SEC filings: Retrieval-augmented generation with XBRL numeric verification. Journal of Advanced Computing Systems, 3(2), 60–76. https://doi.org/10.69987/JACS.2023.30205

Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.19481

Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. Journal of Technology Informatics and Engineering, 4(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces.

International Journal of Graphic Design, 3(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

Zheng, B., Gou, B., Kil, J., Sun, H., & Su, Y. (2024). GPT-4V(ision) is a generalist web agent, if grounded. In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 61349–61385). PMLR. https://proceedings.mlr.press/v235/zheng24e.html

Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. Journal of Advanced Computing Systems, 4(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. Journal of Advanced Computing Systems, 4(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. Journal of Technology Informatics and Engineering, 4(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. International Journal of Graphic Design, 3(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

Zhou, B., Jin, J., & Zhao, D. (2025). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. Journal of Technology Informatics and Engineering, 4(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

Zhou, B., Wang, H., & Chang, X. (2025). Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens. Journal of Technology Informatics and Engineering, 4(2), 447–463. https://doi.org/10.51903/jtie.v4i2.522

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Bisk, Y., Fried, D., Alon, U., & Neubig, G. (2024). WebArena: A realistic web environment for building autonomous agents. International Conference on Learning Representations.