Journal of Knowledge Learning and Science Technology
ISSN: 2959-6386 (Online)
2026, Vol. 5, No. 2, pp. 65–78
DOI: https://doi.org/10.60087/jklst.vol5.n2.005
Journal homepage: https://jklst.org/index.php/home

Uncertainty-Aware Vision–Language Design-to-Code Agents with Iterative Design Critique and Structural Fidelity Evaluation

Elaine Lu

Computer Science, University of Rochester, Rochester, NY,

**Corresponding author: e_lu_rochester@gmail.com*

Abstract

Design-to-code agents must recover visible content, infer latent webpage structure, generate executable HTML and CSS, and recognize when the resulting page is unreliable. This study evaluated an uncertainty-aware agent on all 484 Design2Code webpages. Three GPT-4V conditions—direct generation, text-cue generation, and critique-and-repair—supplied one candidate per page, yielding 1,452 code artifacts and complete browser renders. A CSS-grouped five-fold protocol prevented related style families from crossing fitting and evaluation partitions. Within each fold, a retrieval-render selector ranked the three candidates, and cross-fitted isotonic regression calibrated its confidence for low-confidence clarification. Fidelity was measured through render success, CLIP similarity, rendered-text agreement, block match, position similarity, CIEDE2000 color similarity, and a deterministic DOM composite combining tag overlap, tag-sequence longest-common-subsequence similarity, node-count similarity, and depth similarity. The repair condition gave a language-model design critic the target screenshot, current render, current HTML, and expected visible-text list before one bounded edit. The evaluation compared generation modes, measured iterative gains, and assessed expected calibration error, Brier score, and risk–coverage behavior. The selector achieved extended fidelity of 0.735738, compared with 0.718614 for direct generation, while calibration reduced ECE from 0.592147 to 0.026508. Browser execution was therefore treated as a necessary condition rather than sufficient evidence, connecting multimodal fidelity directly to selection, repair, and clarification decisions.

Keywords: *design-to-code; vision–language models; webpage generation; retrieval-augmented generation; design critique; uncertainty calibration; selective prediction; structural fidelity.*

Article information: Received: 02/03/2026; Accepted: 01/06/2026; Published: 25/06/2026

1. Introduction

A screenshot-to-code system reconstructs an executable program from its rendered surface. The screenshot reveals text, color, geometry, and visual hierarchy, but conceals the DOM, CSS cascade, responsive constraints, and implementation choices that produced them. Different programs can render similarly, while similar markup can produce divergent layouts. The task is therefore an inverse-rendering problem with substantial structural ambiguity, not ordinary image captioning or code completion.

Design2Code established a controlled benchmark of 484 real-world webpages and evaluated generated implementations through browser renders [1]. Its dataset paired source HTML with reference screenshots and first became publicly available in 2024 [2]. Earlier work

demonstrated end-to-end interface-code generation in constrained settings [3], and subsequent resources expanded screenshot–code supervision through synthetic scale, instruction tuning, and real-world layout annotations [4], [5]. Sketch2Code extended the problem to rough sketches and interactive clarification [6], while ScreenAI advanced specialized understanding of UI elements and their locations [7]. These developments improved generation and perception, but reliable automation also requires a policy for accepting, repairing, or deferring a generated page.

We formulated that policy as an evidence-driven agent loop. For each benchmark page, the evaluation paired the reference with direct screenshot-to-code output, output conditioned on extracted text cues, and output produced through one critique-and-repair pass. All 1,452 candidates were parsed and compared with their complete browser screenshots. The repair critic received the target screenshot, current screenshot, current HTML, and expected visible-text list, grounding the edit in observable page content and the program being changed.

Evaluation separated dimensions that a single image score could conflate. CLIP supplied global visual-semantic similarity [8]. Rendered-text agreement compared visible strings; OCR systems such as Tesseract provide relevant context for image-only text recovery [9], but browser-derived text was deterministic for executable artifacts. Tree-edit research established the importance of ordered hierarchy in structural comparison [10]. Motivated by that view, the implemented DOM metric combined tag overlap, preorder tag-sequence alignment, node-count similarity, and maximum-depth similarity. Position and block measures captured layout, while CIEDE2000 represented perceptual color difference [11].

Candidate selection used CSS-grouped five-fold evaluation. Related styling families remained within folds, limiting leakage from near-duplicate designs. Retrieval supplied evidence from analogous training-fold pages, following the separation of parametric generation and non-parametric memory in retrieval-augmented generation [12] and representation-based matching in dense retrieval [13]. The retrieval-render selector then ranked the three held-out candidates from target, render, structural, and neighbor features.

Repair drew on iterative self-feedback [14], verbal reflection [15], and tool-grounded critique [16], but every edit was tested by rendering and scoring the resulting code. Because model-based judging can inherit evaluator biases [17], even when explicit criteria improve human alignment [18], the critic was not used as the numerical judge. Confidence was likewise derived after rendering. Because raw model scores need not represent correctness probabilities [19] ,[20] and can deteriorate under distribution shift [21], selector confidence was calibrated from out-of-fold predictions with isotonic regression. A low-confidence result triggered clarification, operationalizing the risk–coverage principle of selective classification [22], integrated rejection [23] , and the broader uncertainty-control perspective of conformal prediction [24].

This study contributes a paired evaluation over 484 pages and 1,452 candidates; a leakage-conscious retrieval-render selector; a fidelity suite spanning execution, appearance, text, layout, color, and DOM structure; a screenshot-and-code-grounded repair condition; and cross-fitted confidence linked to clarification. Together, these components frame design-to-code as selective multimodal engineering in which an agent generates code, observes its browser consequences, repairs concrete defects, and justifies automatic delivery.

2. Literature Review

Early screenshot-to-code research concentrated on direct mapping from pixels to a constrained interface language. Pix2code showed that an encoder-decoder could generate representations for web, Android, and iOS interfaces [3]. WebSight later supplied millions of synthetic HTML–screenshot pairs for vision–language training [4]. Web2Code combined webpage generation with instruction and question-answer data, linking code production to page understanding [25]. WebCode2M emphasized large-scale real-world examples, layout annotations, and hierarchical structure [5]. These datasets supported model development, whereas Design2Code supplied a compact real-world evaluation set with rendered references, fine-grained measures, and human validation [1], [2].

Input ambiguity motivated interactive formulations. Sketch2Code evaluated hand-drawn sketches and introduced feedback-following and question-asking interactions [6]. ScreenAI used screen annotations and UI-focused vision–language training to represent element type and location [7]. Both works indicate that visual grounding alone does not resolve every design choice. The clarification mechanism studied here addresses the same practical issue through calibrated selection confidence: interaction is requested when measured evidence does not support automatic acceptance.

Fidelity evaluation requires complementary representations. CLIP embeddings capture global visual semantics but may underweight exact wording, small displacement, and hierarchy [8] , documented a mature OCR architecture for recovering text from images [9]; when executable webpages are available, browser-extracted visible text supports agreement measurement without recognition noise. Tree edit distance formalizes structural difference between ordered labeled trees [10]. The DOM composite adopted its hierarchical motivation while using four transparent components—tag overlap, tag-sequence LCS, node count, and depth—that remain stable and directly interpretable. CIEDE2000 supplies a perceptually grounded color-difference formulation [11]. Render success, block match, and position similarity complete the metric suite.

Learned evaluators can diagnose broad design qualities, but their outputs require care. LLM-as-a-judge studies reported scalable evaluation together with position, verbosity, and self-preference biases [17]. G-Eval improved structured model-based judgment through explicit criteria and reasoning procedures [18]. Accordingly, the language-model critic served as a repair planner, not the quantitative judge. It inspected paired screenshots, current HTML, and the visible-text list; the repaired page was then rerendered and scored with deterministic and embedding-based measures.

Retrieval and iterative correction supply two forms of external grounding. RAG combines parametric generation with non-parametric evidence [12], and dense passage retrieval demonstrated representation-based retrieval beyond lexical matching [13]. Transferred to webpage artifacts, these ideas support retrieval from stylistically related training-fold pages after candidate rendering. Self-Refine showed that generation can be followed by feedback and refinement without parameter updates [14]. Reflexion used verbal feedback to improve subsequent agent behavior [15] while CRITIC coupled correction to external tools [16] ReAct combined reasoning traces with tool-mediated actions [26]. These approaches motivate a

bounded render-and-repair loop in which the critic receives concrete visual and program context.

Calibration connects fidelity estimation to action. Guo et al. showed that neural confidence can be systematically miscalibrated [19]. Mukhoti et al. showed that the training objective can materially improve calibration [20], and Ovadia et al. demonstrated additional degradation under distribution shift [21]. Selective classification formalizes the trade-off between accepted coverage and error among accepted cases [22]. SelectiveNet integrates rejection into model training [23], while conformal prediction provides a broader framework for distribution-free uncertainty control under its validity conditions [24]. The selector used cross-fitted isotonic regression rather than an integrated reject head. Its calibrated probability supports ECE, Brier, and risk–coverage analysis and converts low confidence into a focused clarification decision.

Recent design-generation work expands this foundation from code output to explicit design reasoning. Chen and Li evaluated sketch-to-prototype conversion with structural and visual consistency [27], complementing interactive Sketch2Code [6]. Human alignment is addressed by an LLM design critic whose feedback is compared with graphic-design judgment [28]. Text-grounded rationale cards translate advertising-layout metadata into inspectable decisions [29]. Designer-editable briefs [30] and structured intention-to-decision cards [31], preserve revision control. These studies motivate a critic that identifies observable defects and proposes bounded edits instead of silently replacing a candidate.

Interface evidence can also be behavioral or system-level. Review and screen analysis informed adaptive volleyball-learning interfaces [32], while dark-pattern detection treated interface microcopy as semantically consequential design evidence [33]. Edge quality prediction via VMAF distillation converts perceptual quality into compact tokens [34], and layout-aware PDF slicing prioritizes content that makes a document functionally usable sooner [35]. For webpage reconstruction, these findings support separating appearance, wording, layout, execution, and structure rather than treating visual similarity as a complete quality judgment.

Evidence-card systems show how trustworthy interfaces can expose sources and uncertainty. Scientific-search cards organize retrieval transparency and visual evidence hierarchy [36]; accounting disclosure cards anchor explanations to financial statements and notes [37] . Recommendation cards communicate product rationale under calibrated trust [38], and financial dashboards study hierarchy and chart comprehension for institutional risk [39]. Medical image explanation cards bind visual evidence to uncertainty [40], while breast-ultrasound cards specialize that relationship for image-based diagnostic support [41]. These patterns inform the paper's use of visible mismatch evidence and calibrated confidence as distinct review signals.

Risk-communication research further separates confidence from action. Patient-facing safety cards translate a rubric into visible medical-risk cues [42], and a related benchmark evaluates explainable response interfaces [43]. Counseling-support cards attach ranked evidence to intake recommendations [44]. Privacy and data-integrity cards expose prompt-injection hazards to human reviewers [45], incident cards convert cloud logs into actionable evidence [46], capacity briefs turn workload risk into dashboard decisions [47], and teacher-facing education explanations show how prediction outputs can be organized for intervention [48]. The domains

differ, but the interface requirement is consistent: evidence, risk, and the next action should remain visually distinguishable.

Uncertainty-aware systems reinforce that confidence should be aligned to decision risk, not to the largest raw score. Late fusion calibrates component confidence before learning a fusion rule [49]; resume-job screening couples pairwise matching to selective refusal [50]. Lightweight biomedical vision-language classification makes uncertainty an explicit evaluation target [51], while cross-modal medical pretraining studies parameter-efficient few-shot transfer [52]. GPU-demand forecasting reports operational risk curves across horizons [53]. These studies support the paper's separation of candidate generation, post-render quality prediction, probability calibration, and selective action.

Generalization under limited or shifted evidence adds a second constraint. Few-shot cold-start workload forecasting uses time-series foundation models for unseen inference tenants [54]. Evidence-calibrated RAG measures retrieval coverage, abstention, and hallucination proxies on unanswerable questions [55]. Object-level LiDAR-camera tracking preserves modality-specific evidence before probabilistic association [56], and trust-calibrated multilingual RAG evaluates whether confidence remains meaningful across languages [57]. Although their tasks differ, all four treat transfer and uncertainty as properties to be measured rather than assumed.

Risk-sensitive explanations provide complementary controls. Fair credit scoring couples predictive output to counterfactual recourse [58]. The uncertainty of spectral estimators and maximum-likelihood synchronization methods has been analyzed directly [59]. Conformal alert control connects anomaly evidence to bounded incident-ticket generation [60], while domain adaptation with approximately shared features formalizes transfer when source and target representations overlap only partially [61]. These results motivate grouped out-of-fold evaluation, calibrated acceptance, and explicit discussion of distribution shift in design-to-code deployment.

Retrieval reliability literature narrows grounding from simply using documents to controlling the evidence path. Budgeted multi-hop retrieval evaluates which hops justify additional cost [62], and evidence-chain RAG requires source attribution and deterministic provenance [63]. Scientific-claim verification ranks evidence against narrative context [64]; accounting-aware due diligence constrains retrieval to disclosures [65]. Offline financial search reranking combines query rewriting, hybrid retrieval, and listwise relevance ranking [66]. For webpage routing, these approaches suggest treating retrieved neighbors as auditable decision evidence rather than as implicit permission to copy their code.

Other evidence-constrained agents combine retrieval with deterministic checks. Machine-learning prediction of RAG quality estimates when retrieved context is weak before generation [67]. A therapist-facing copilot retrieves and ranks support from multi-turn dialogue [68], while numerical guardrails validate quantitative answers over SEC and FRED data [69]. Tokenized-asset risk monitoring constrains disclosure and settlement claims to accounting evidence [70]. Retrieval-grounded HDFS diagnosis links anomaly decisions to failure narratives [71]. In the present setting, browser execution and metric decomposition play the same verification role: retrieved similarity can influence routing, but observed artifact quality remains decisive.

Faithfulness also matters when retrieved evidence drives personalized or regulated outputs. Review-grounded recommendation evaluates whether explanations are supported by product reviews [72]; behavior retrieval plus response generation makes conversational recommendations interpretable [73]. Multi-regulation RAG structures legal evidence for cross-border product counsel [74]. Retrieved normal prototypes provide concrete comparators for OpenStack failure explanations [75]. These works reinforce the distinction between generation evidence, selector evidence, and the final explanation shown to a reviewer.

Routing and resource studies address how confidence becomes operational policy. Conservative off-policy selection avoids choosing brittle bidding and ranking policies [76]. Hybrid-cloud LLM deployment treats cost and placement as part of model serving [77]. Profit-aware GPU admission combines cost-sensitive prediction with evidence-grounded policy memos [78], and power-aware inventory planning joins job-level forecasts to workload explanations [79]. Their common contribution is decision-aware evaluation: a model is judged not only by average prediction error but also by consequences of the action it triggers.

Reliability predictors and guarded workflows extend that view. Generalist-agent trajectory models forecast tool-use failure before task completion [80]. Cost-aware AIOps routing selects among parsing, anomaly detection, diagnosis, and summarization paths [81]. GPU-memory prediction informs whether a deep-learning task can be admitted safely [82], and CloudTrail sequence models identify interpretable attack-chain stages with temporal smoothing [83]. These operational analogies motivate the selector's leakage-controlled training and a clarification route when predicted fidelity is insufficient.

Human oversight also depends on privacy, rubrics, response control, and stable ranking. Differentially private topic-preference learning limits exposure in knowledge-grounded chat [84]. Rubric-grounded essay scoring makes automated feedback auditable [85], while strategy-aware therapist imitation controls the response pattern used in emotional-support dialogue [86]. Spectral rank-aggregation theory examines when pairwise preferences yield a reliable ordering [87]. Early-warning education analytics pairs risk predictions with intervention rationales [88]. In a design-to-code agent, the corresponding requirements are inspectable criteria, stable candidate ordering, minimal disclosure of user content, and a clear handoff when automatic completion is not justified.

3. Methodology

The 77,578,732-byte Design2Code Parquet corpus contained 484 records, each comprising screenshot bytes and source HTML. Content hashing aligned all records with stable page identifiers, with no missing record, corrupt screenshot, or unmatched row. Screenshot height ranged from 720 to 3,328 pixels; median HTML length, DOM size, DOM depth, and visible-text length were 63,048 characters, 119 elements, 10 levels, and 1,357.5 characters, respectively. Images occurred on 364 pages, forms on 205, and tables on 86. Table 1 reports the complete data profile.

Table 1. Design2Code data profile and integrity checks

Characteristic	Minimum / count	Median	Maximum / count
Pages	484	484	484
Screenshot height (px)	720	1,351	3,328
HTML characters	1,744	63,048	241,096
DOM elements	8	119	508
DOM depth	3	10	31
Visible text characters	32	1,358	3,339
Missing screenshot or HTML	0	—	0
Corrupt PNG files	0	—	0
Exact duplicate pairs	0	—	0
Recoverably parsed HTML pages	484	—	484

The source file contained 484 screenshot-HTML pairs. Element counts exclude comments; screenshot dimensions were read from the embedded PNG byte streams.

The Parquet schema stored each screenshot as binary PNG data and each implementation as a text field. Extraction preserved the original bytes, computed screenshot and HTML SHA-256 digests, and wrote a row-to-page map before any metric was calculated. PNG verification, nonempty-HTML checks, and exact pair hashing found no corrupt or duplicated record. HTML was parsed with recovery enabled; DOM counts and depth were measured from the body subtree, while script, style, template, and noscript content was excluded from visible-text statistics. These decisions kept the corpus audit deterministic across malformed but browser-recoverable markup.

Every source page was paired with the released GPT-4V direct, text-augmented, and visual-repair artifacts, yielding 1,452 candidate HTML–screenshot pairs. All candidates had nonempty HTML and valid rendered images. Table 2 defines the three fixed candidate paths, the retrieval-render selector, and the calibrated confidence gate. A per-page upper bound, reported with the policy results, selected the best of the three GPT-4V candidates after observing their outcomes and served only as a diagnostic ceiling. The selector instead predicted candidate quality from available screenshot, render, retrieval, and candidate-structure evidence. Low-confidence clarification was implemented through selective abstention: pages retained by a confidence threshold were auto-completed, while the remainder were sent for clarification or manual review. This controller follows the same decision-aware separation found in selective refusal [50], evidence-calibrated abstention [55], conservative policy selection [76], trajectory reliability forecasting [80], and cost-aware workflow routing [81].

Table 2. Evaluated generation and routing policies

Policy	Evidence	Action	Role
Direct VLM	Target screenshot	Single-pass HTML/CSS generation	Fixed baseline
Text cue	Screenshot plus complete visible-text list	Content-grounded generation	Fixed text-aware policy
Critique-repair	Target render, current render, HTML, and text cue	One design-critic repair pass	Fixed repair policy

Policy	Evidence	Action	Role
Retrieval-render selector	Target features, 10 neighbors, and candidate-render evidence	Selects one of three candidates	Grouped out-of-fold policy
Calibrated confidence gate	Cross-fitted selector score	Auto-completes or requests clarification	Selective prediction

The three candidate families were the publicly released GPT-4V artifacts. The routing and confidence analyses were executed on all 484 matched pages.

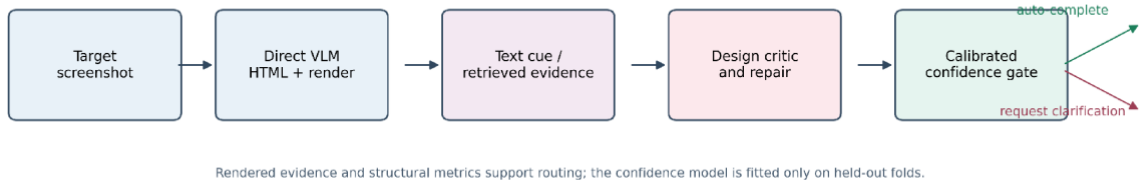

Figure 1. Uncertainty-aware design-to-code workflow. Candidate generation, retrieved evidence, render-based critique, and calibrated gating form one execution path.

The official benchmark metrics were block match, text match, normalized position similarity, perceptual color similarity, and CLIP ViT-B/32 screenshot similarity. Their unweighted mean defined official fidelity.

$$F_{official} = (F_{block} + F_{text} + F_{position} + F_{color} + F_{CLIP}) / 5$$

DOM fidelity combined tag-multiset F1, normalized longest-common-subsequence similarity over preorder tag sequences, node-count similarity, and tree-depth similarity.

$$F_{DOM} = 0.35F_{tag} + 0.35F_{LCS} + 0.15F_{nodes} + 0.15F_{depth}$$

The primary outcome added an explicit structural term to the official benchmark score.

$$F_{extended} = 0.75F_{official} + 0.25F_{DOM}$$

Success required extended fidelity of at least 0.80. HTML token F1, normalized pixel similarity, 16-bin RGB histogram intersection, edge cosine, and global SSIM were retained as diagnostics. Table 3 defines all metrics and separates the official five-metric score from extended fidelity.

Table 3. Evaluation metrics and operational definitions

Metric	Definition	Scale
Block match	Area-weighted recall and hallucination penalty for matched visual blocks	0-1; higher is better
Text match	Character-level similarity over matched rendered text blocks	0-1; higher is better
Position	Normalized center-coordinate agreement for matched blocks	0-1; higher is better
Color	Perceptual text-color agreement based on CIEDE2000	0-1; higher is better

Metric	Definition	Scale
CLIP	Cosine agreement of masked screenshot embeddings	0-1; higher is better
DOM fidelity	0.35 tag F1 + 0.35 preorder LCS + 0.15 node-count + 0.15 depth similarity	0-1; higher is better
Render success	Nonempty HTML with a decodable generated PNG	Binary
Extended fidelity	0.75 mean of five visual metrics + 0.25 DOM fidelity	0-1; higher is better
ECE / Brier / NLL	Reliability and proper-scoring diagnostics for success probability	Lower is better
AURC	Area under selective risk versus auto-completion coverage	Lower is better

Success for calibration was defined before fitting as extended fidelity ≥ 0.80 with a valid render artifact.

Official per-page scores were joined by stable page identifier and method name. DOM and image diagnostics were recomputed from the matched reference and candidate files. Pixel comparison used aspect-preserving white canvases so that pages with different heights remained comparable, and render success was verified independently of fidelity. Diagnostic pixel, histogram, edge, and SSIM values were not added to the primary outcome; they supported the selector and helped localize failure without changing the declared weighting of visual and structural fidelity.

Each target screenshot was aspect-preservingly fitted to a white 20-by-20 canvas and represented by flattened RGB values, 12-bin channel histograms, grayscale mean and standard deviation, gradient mean and standard deviation, and the 0.1, 0.5, and 0.9 grayscale quantiles. Fold-specific principal-component analysis retained at most 64 components, followed by vector normalization and cosine retrieval.

Normalized style contents produced 472 CSS groups: 460 single-page groups and 12 two-page groups. Five-fold GroupKFold kept every CSS group wholly within one fold. For each test page, ten training neighbors supplied method-specific quality evidence weighted by $1/(d + 0.05)$; leave-one-out neighbors produced the training counterpart. Candidate features comprised CLIP, pixel similarity, histogram intersection, edge cosine, global SSIM, candidate DOM size and depth, render validity, method identity, and retrieved neighbor quality. Reference-DOM fidelity and the extended-fidelity target were excluded from selector inputs. Keeping these evidence families explicit is consistent with inspectable design rationales [29], compact quality representations [34], and visually ordered evidence cards [36], rather than an opaque single-score judgment.

A GradientBoostingRegressor predicted extended fidelity using 120 estimators, learning rate 0.035, maximum depth 2, Huber loss, and seed 20260729. The candidate with the highest prediction was selected. Every reported policy result was out of fold. Isotonic regression then mapped predicted quality to the binary success event through a second grouped five-fold cross-fitting procedure.

The controller was fitted at candidate level, giving three training rows per reference page, but selection and reporting remained page-level. Each test page was scored only by a model whose fitting set excluded that page and every page sharing its CSS signature. Reference HTML,

reference DOM size, and realized extended fidelity never entered the controller feature vector. The target screenshot was available because it was the design specification, and candidate HTML and renders were available because selection occurred after generation and browser execution.

Calibration was evaluated with ten equal-frequency-bin ECE, Brier score, and negative log-likelihood. Risk–coverage curves ranked pages by calibrated probability and measured risk as one minus accepted-page mean extended fidelity at 19 coverage levels from 10% to 100%. Reference DOM size defined four complexity groups containing 125, 118, 120, and 121 pages. Means used 4,000-resample bootstrap intervals. Selector comparisons used paired Wilcoxon signed-rank tests with Pratt zero handling and Holm adjustment. Failure categories were assigned by the minimum component among block, text, position, color, CLIP, and DOM fidelity. The rank-calibrate-act sequence parallels calibrated fusion [49], multilingual trust calibration [57], conformal alert control [60], and reliability forecasting for tool-using agents [80].

4. Results

All 1,452 GPT-4V candidates passed the render-validity test. Figure 2 compares page 11699, which had the largest critique-repair gain over direct generation, with page 4300, located at the median gain. On page 11699, extended fidelity increased from 0.107462 to 0.765104; on page 4300, it increased from 0.742562 to 0.750346.

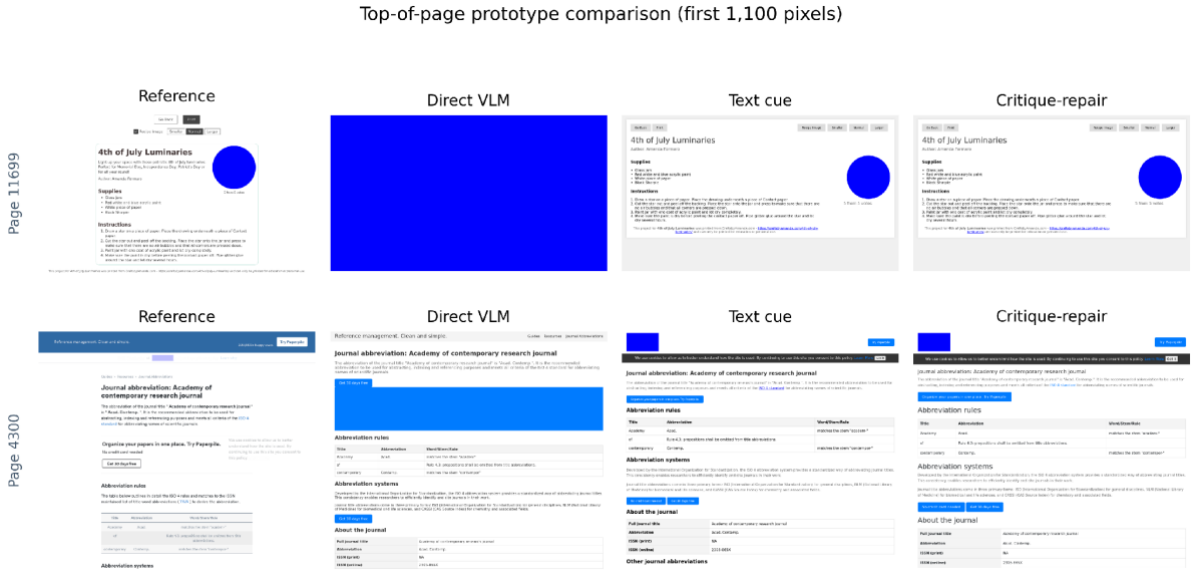

Figure 2. Prototype comparison for pages 11699 and 4300. The display shows the first 1,100 pixels of the reference and three candidate renders.

The eight released generation methods are compared in Table 4. GPT-4V critique-repair ranked first in official fidelity at 0.856270 (95% CI 0.851404–0.861000), followed by GPT-4V text augmentation at 0.852371 and direct prompting at 0.847789. The remaining means were 0.804929 for Gemini text augmentation, 0.803927 for Design2Code-18B-v0, 0.801417 for Gemini visual repair, 0.794447 for Gemini direct prompting, and 0.771211 for WebSight VLM.

Figure 3 shows that critique-repair's lead came mainly from block, position, and DOM fidelity rather than uniform gains across every component.

Table 4. Full-benchmark five-metric performance across eight released methods

Method	Block	Text	Position	Color	CLIP	Mean
GPT-4V critique-repair	0.888	0.981	0.811	0.729	0.872	0.856
GPT-4V text cue	0.876	0.982	0.802	0.730	0.872	0.852
GPT-4V direct	0.858	0.974	0.805	0.733	0.869	0.848
Gemini text cue	0.848	0.969	0.704	0.663	0.840	0.805
Design2Code-18B	0.785	0.964	0.743	0.670	0.858	0.804
Gemini critique-repair	0.841	0.966	0.701	0.662	0.837	0.801
Gemini direct	0.802	0.946	0.723	0.662	0.839	0.794
WebSight VLM	0.559	0.866	0.773	0.794	0.865	0.771

Each value is the mean across 484 pages. Mean is the unweighted average of block, text, position, color, and CLIP fidelity.

The component profile clarifies the ranking. GPT-4V critique-repair reached block, text, position, color, and CLIP means of 0.887954, 0.980731, 0.811199, 0.728978, and 0.872486. Text-cue generation obtained the highest text mean, 0.981616, whereas direct generation obtained the highest color mean, 0.732987. WebSight VLM's official mean was limited mainly by block match at 0.558970, despite CLIP similarity of 0.864549. Thus, aggregate order reflected distinct error mixtures rather than one method dominating every dimension.

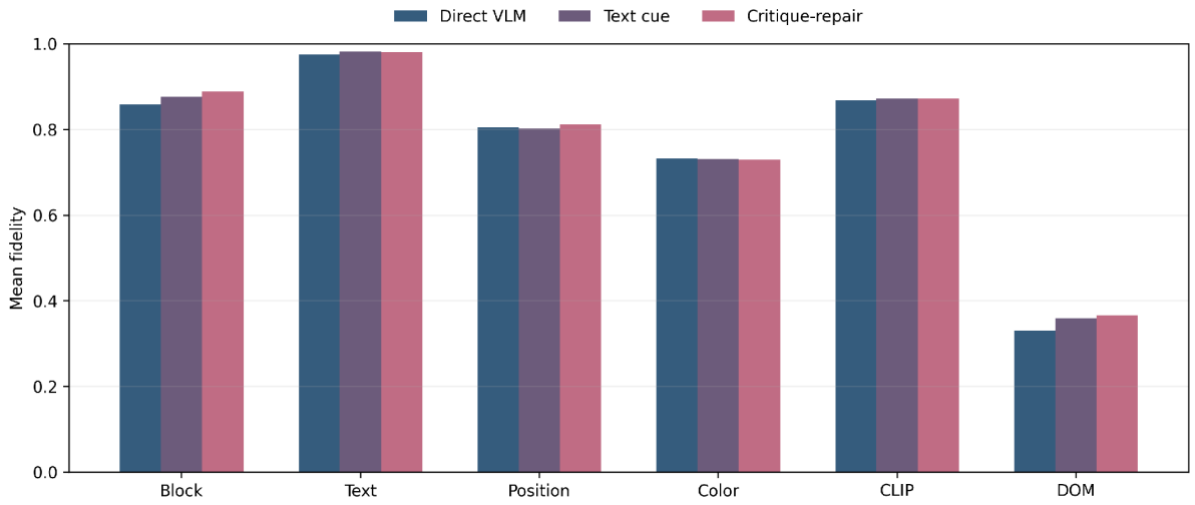

Figure 3. Mean block, text, position, color, CLIP, and DOM fidelity for the three GPT-4V candidate families.

Table 5 gives the structural and render results for the three GPT-4V variants. Direct generation achieved official fidelity 0.847789, DOM fidelity 0.331090, and extended fidelity 0.718614. Text augmentation raised these values to 0.852371, 0.358433, and 0.728886. Critique-repair reached 0.856270, 0.365784, and 0.733648, while also leading in HTML text F1 (0.873183), pixel similarity (0.805971), histogram intersection (0.775160), and global SSIM (0.188463).

Table 5. Structural, textual, pixel, and render results for GPT-4V candidates

Variant	Visual mean	Tag F1	DOM LCS	DOM composite	HTML text F1	Pixel sim.	Render
Direct VLM	0.848	0.368	0.235	0.331	0.797	0.794	100.0%
Text cue	0.852	0.398	0.261	0.358	0.847	0.800	100.0%
Critique-repair	0.856	0.407	0.269	0.366	0.873	0.806	100.0%

All 1,452 candidate HTML files had a corresponding decodable render. DOM metrics compare each candidate with its matched reference HTML.

Bootstrap intervals supported the same ordering. Direct extended fidelity had a 95% interval of 0.711317–0.725515, text cue 0.722611–0.735121, and critique-repair 0.727800–0.739575. The much lower DOM means than official visual means exposed a systematic gap: critique-repair's visual mean exceeded its DOM composite by 0.490486. Candidate pages could therefore preserve text and overall appearance while simplifying the reference hierarchy, which motivated retaining DOM fidelity as a separate quarter of the primary score.

Structural fidelity declined sharply with complexity. Figure 4 shows critique-repair DOM fidelity falling from 0.526725 in the lowest-complexity group to 0.218209 in the highest; direct generation fell from 0.489177 to 0.187602. Figure 5 further shows heterogeneous iteration effects. Text augmentation improved on direct generation for 55.58% of pages with mean gain 0.010272. Critique-repair improved on direct generation for 58.68%, averaging 0.015034, but page-level differences ranged from −0.157871 to 0.657642.

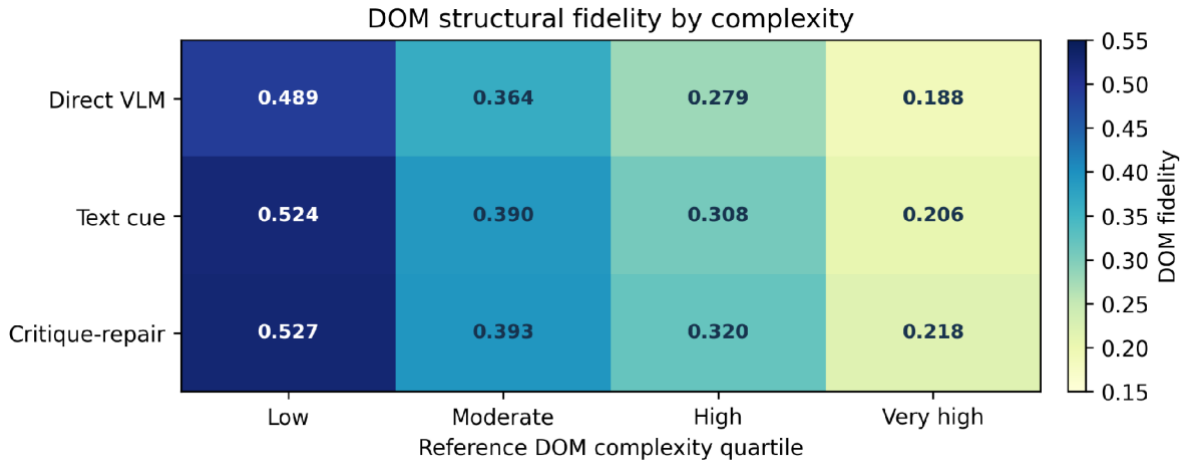

Figure 4. DOM structural fidelity by reference-page complexity. Structure weakened monotonically as reference DOM size increased.

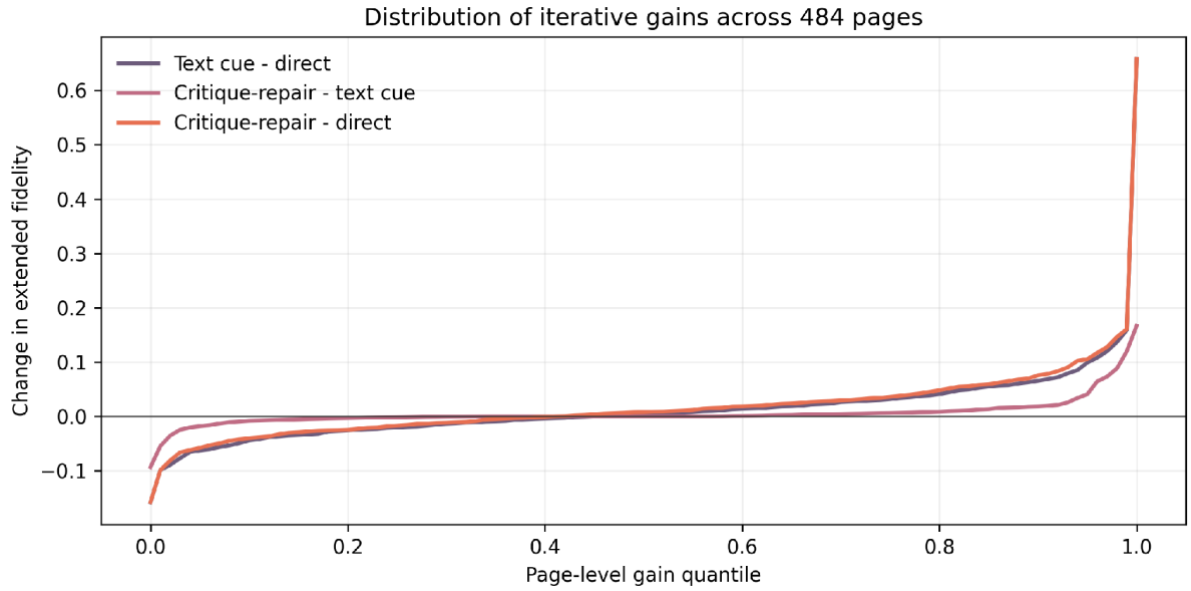

Figure 5. Page-level distribution of gains from text cues and one critique-repair pass. Positive values indicate improvement over the preceding or direct stage.

Repair was therefore neither uniformly beneficial nor merely cosmetic. Relative to the text-cue condition, it improved 44.83% of pages, tied on 18.80%, and produced mean gain 0.004762. These mixed outcomes created the selection opportunity: a universal repair policy discarded a stronger earlier candidate on some pages, whereas the selector could preserve direct or text-cue output when its post-render evidence predicted higher fidelity.

The selector routed 216 pages (44.63%) to critique-repair, 167 (34.50%) to direct generation, and 101 (20.87%) to text augmentation. Table 6 reports selector extended fidelity of 0.735738 (95% CI 0.729985–0.741314) and a 14.67% success rate. The per-page upper bound reached 0.747153 and 17.98%, so the selector recovered 60.00% of the direct-to-upper-bound gap.

Table 6. Policy fidelity, success rate, and selector routing

Policy	Extended fidelity	95% bootstrap CI	Success	Pages selected	Routing share
Direct VLM	0.718614	[0.711510, 0.725616]	11.57%	167	34.50%
Text cue	0.728886	[0.722915, 0.735136]	14.67%	101	20.87%
Critique-repair	0.733648	[0.727634, 0.739446]	14.05%	216	44.63%
Retrieval-render selector	0.735738	[0.729985, 0.741314]	14.67%	—	—
Per-page upper bound	0.747153	[0.741720, 0.752568]	17.98%	—	—

Routing counts apply to the retrieval-render selector: direct 167 pages, text cue 101 pages, and critique-repair 216 pages.

Paired results in Table 7 show a selector gain over direct generation of 0.017124 (95% CI 0.012610–0.022347; Holm-adjusted $p = 4.77 \times 10^{-13}$). The gain over text augmentation was 0.006852 (0.004283–0.009432; $p = 8.09 \times 10^{-8}$). Against critique-repair, the mean was 0.002090; its bootstrap interval crossed zero (−0.000432–0.004547), although the adjusted rank-test result was $p = 0.010619$, indicating a small heterogeneous advantage.

Table 7. Paired selector comparisons over the same 484 pages

Baseline	Mean gain	95% bootstrap CI	Holm p	Improved	Tied
Direct VLM	0.017124	[0.012610, 0.022347]	4.77e-13	44.63%	34.50%
Text cue	0.006852	[0.004283, 0.009432]	8.09e-08	42.77%	32.85%
Critique-repair	0.002090	[-0.000432, 0.004547]	0.0106	31.82%	44.83%

Two-sided Wilcoxon signed-rank tests used Pratt zero handling; p-values were adjusted across the three comparisons by Holm's procedure.

Table 8 shows selector fidelity of 0.796052, 0.746418, 0.718481, and 0.680129 across increasing complexity, compared with direct-generation values of 0.782944, 0.724682, 0.704072, and 0.660661. Routing therefore retained gains at every complexity level.

Table 8. Extended fidelity by reference DOM-complexity quartile

Complexity	Direct VLM	Text cue	Critique-repair	Retrieval-render selector
Low	0.782944	0.796029	0.795218	0.796052
Moderate	0.724682	0.737582	0.741134	0.746418
High	0.704072	0.710156	0.718949	0.718481
Very high	0.660661	0.669618	0.677321	0.680129

Quartiles were formed from reference DOM element counts. Group sizes were 125, 118, 120, and 121 pages.

Table 9 reports that cross-fitted isotonic calibration reduced ECE from 0.592147 to 0.026508, Brier score from 0.468951 to 0.101458, and negative log-likelihood from 1.174101 to 0.385869. Figure 6 confirms substantially closer agreement between predicted and observed success. At approximately 10% coverage, the calibrated selector achieved mean fidelity 0.794834 versus 0.785779 for the direct-confidence gate; at 50%, the values were 0.754442 and 0.738489. Figure 7 shows the full trade-off, with selector AURC 0.219811 versus 0.233123.

Table 9. Calibration and selective-risk diagnostics

Assessment	ECE	Brier	NLL	AURC	Risk @20%	Risk @50%	Risk @100%
Raw controller score	0.592147	0.468951	1.174101	—	—	—	—
Cross-fitted isotonic	0.026508	0.101458	0.385869	—	—	—	—
Direct-confidence gate	—	—	—	0.233123	0.232239	0.261511	0.281386
Calibrated retrieval-render selector	—	—	—	0.219811	0.215996	0.245558	0.264262

ECE used ten equal-frequency bins. Calibration probabilities were produced by five-fold cross-fitted isotonic regression; lower values indicate better reliability or ranking.

In the highest equal-frequency bin, the raw controller assigned mean confidence 0.788900 while the observed success rate was 0.604167. After cross-fitted isotonic calibration, the highest bin had mean probability 0.618055 and observed rate 0.562500. Calibration did not alter the

14.67% prevalence or the chosen candidate; it changed the confidence scale used for acceptance. The 5.71% AURC reduction shows that selector confidence retained higher-fidelity pages earlier than the direct-confidence gate as automatic coverage increased.

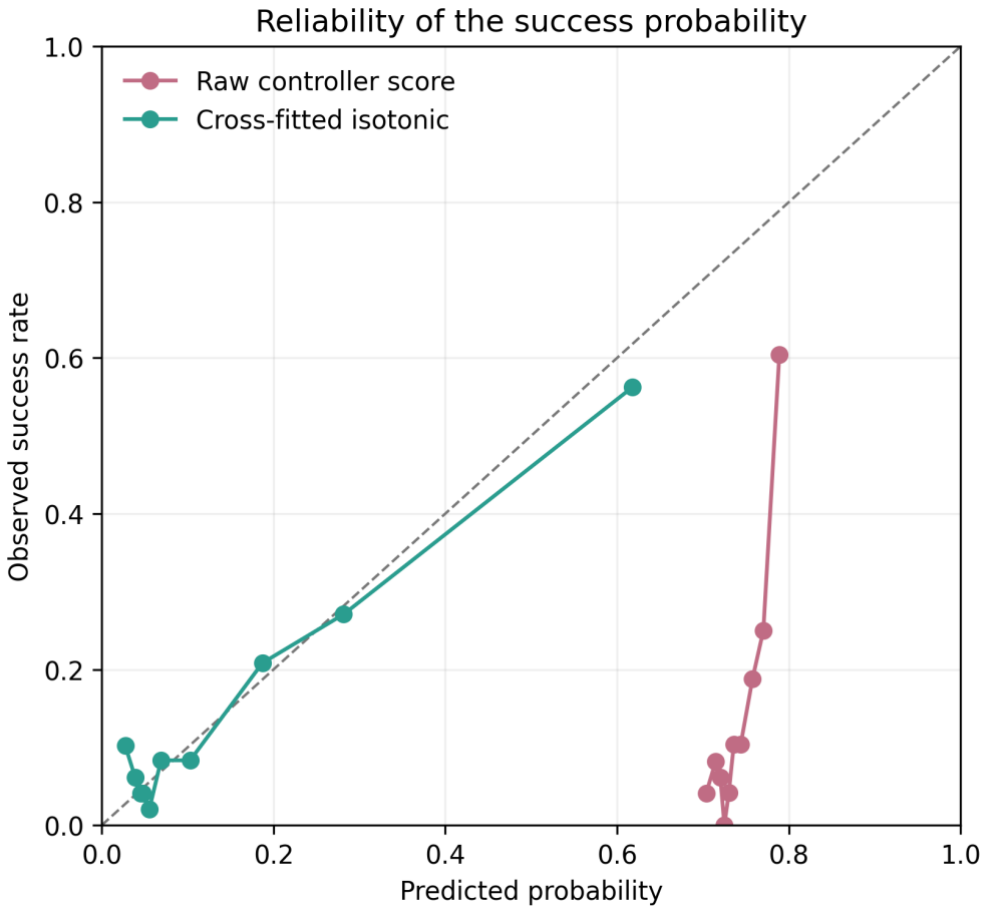

Figure 6. Reliability diagram for the selector success probability before and after cross-fitted isotonic calibration.

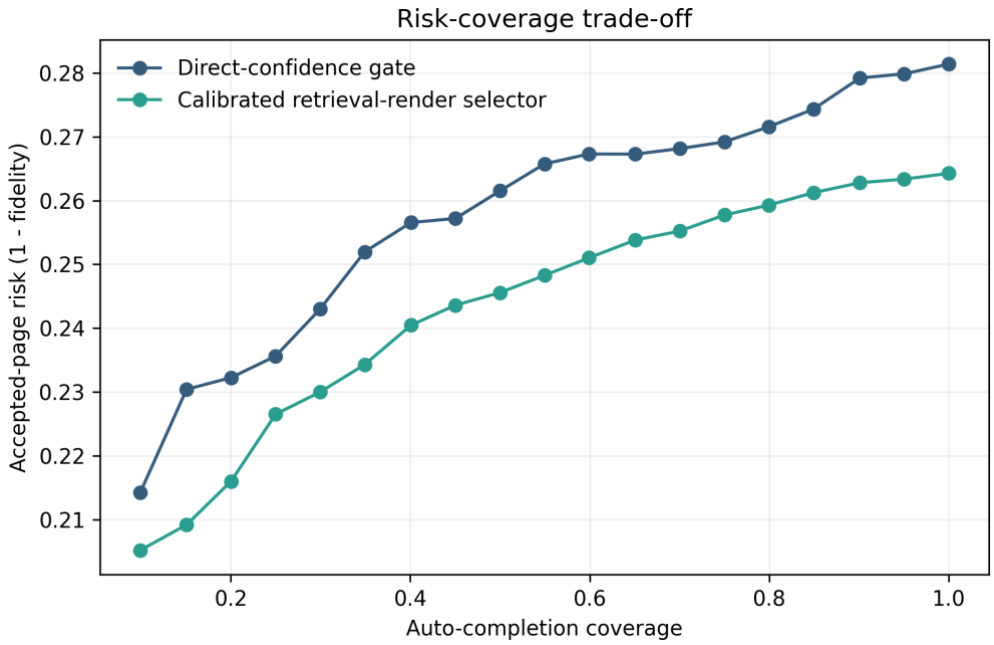

Figure 7. Selective risk as auto-completion coverage increased. The calibrated retrieval-render selector retained lower risk than the direct-confidence gate.

Of 413 unsuccessful selections, Table 10 attributes 399 (96.61%) to DOM fidelity, eight (1.94%) to color, and six (1.45%) to block match. Figure 8 localizes page 11699's remaining spatial discrepancies: despite pixel similarity 0.943533, its DOM fidelity was 0.498755, confirming that strong surface correspondence did not guarantee structural success.

Table 10. Dominant deficit among selector failures

Dominant deficit	Pages	Share of 413 failures	Mean weakest component
DOM structure	399	96.61%	0.309871
Color	8	1.94%	0.298857
Block match	6	1.45%	0.194722

A failure had extended fidelity below 0.80. The dominant deficit was the lowest of block, text, position, color, CLIP, and DOM fidelity.

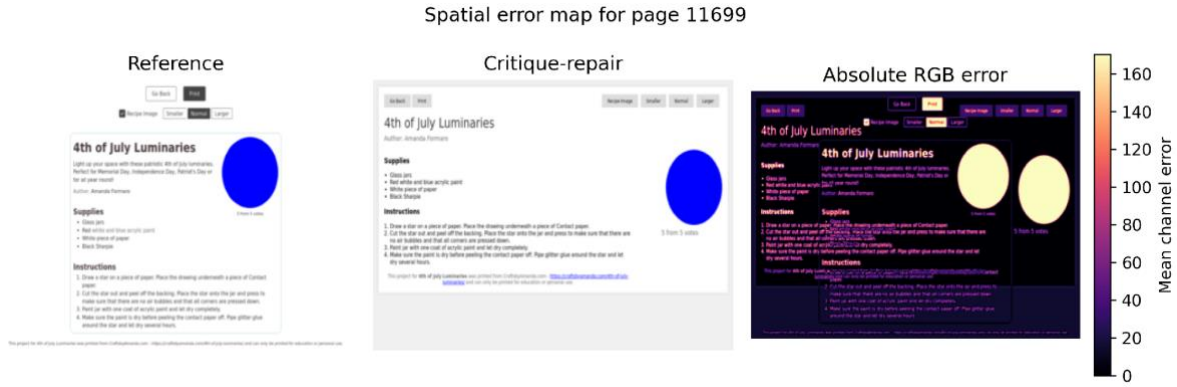

Figure 8. Spatial error map for page 11699 after critique-repair. Bright regions identify large absolute RGB disagreement.

5. Discussion

The retrieval-render selector achieved the highest mean extended fidelity, outperforming every fixed GPT-4V policy. Paired tests showed strong, Holm-adjusted significance against the direct and text-cue methods. Its advantage over critique-repair was smaller: the Wilcoxon test remained significant, but the confidence interval for the mean difference crossed zero. This result supports conditional routing while indicating that critique-repair already captures much of the attainable improvement. It also aligns with human-evaluated design criticism [28] , designer-editable revision artifacts [30], conservative policy selection [76], and guarded workflow routing [81], all of which preserve alternatives until evidence supports a particular action.

Critique-repair was selected for 44.63% of pages, compared with 34.50% for direct generation and 20.87% for text-cue generation. Critique was therefore central, but invoking it universally was less effective than preserving a stronger first-pass candidate when repair was unlikely to help. The remaining gap between the selector and the per-page upper bound also indicates that further gains depend partly on more accurate routing.

Page complexity remained the principal challenge. Selector fidelity declined from 0.796052 for low-complexity pages to 0.680129 for very-high-complexity pages, and every fixed policy showed the same downward pattern. Routing reduced the penalty but could not eliminate the difficulty introduced by larger and deeper page structures. The failure taxonomy explains this result: DOM fidelity was the dominant deficit in 399 of 413 failures, whereas color and block matching dominated only a small minority. Strong visual or textual resemblance therefore did not guarantee faithful reconstruction of element hierarchy, order, and depth. This separation echoes structural-and-visual prototype evaluation [27], layout-prioritized rendering [35], uncertainty-linked visual explanation [40], and incident-card designs that preserve an evidence hierarchy [46]. Future repair mechanisms should optimize DOM structure alongside rendered appearance.

Calibration made the routing scores substantially more useful. Cross-fitted isotonic calibration reduced ECE and Brier score by large margins, while the selector's lower AURC indicated better ranking of pages by expected risk. These calibrated scores can support acceptance, rerouting, or abstention decisions. Comparable action-oriented uses of uncertainty appear in calibrated late fusion [49], selective resume screening [50], unanswerable-question RAG [55], multilingual information access [57], and conformal incident alerting [60]. The selective gate, however, evaluates routing and abstention rather than the quality of a later human answer or its effect on the generated page.

The results also separate the roles of retrieval and critique. Retrieval did not supply code or copy an analogous DOM into the output; it supplied empirical evidence about which candidate family tended to work on nearby designs. Critique produced a new candidate, but its value was established only after browser execution. This evidence boundary is consistent with budgeted retrieval [62], deterministic provenance chains [63], hybrid reranking [66], retrieval-quality prediction [67], grounded failure narratives [71], and prototype-based anomaly explanations [75]. The division makes the system easier to audit: generator quality is measured within each

path, routing quality is measured against fixed page-level outcomes, and calibration quality is measured against held-out success events.

The findings are bounded by the evaluated setting. The candidate generations incorporated source-extracted text cues, so performance does not represent a pure screenshot-only workflow. Fixed released artifacts preserve a stable comparison but do not capture generation variability or newer model capabilities. The outputs were static HTML/CSS; interaction logic, responsive behavior, dynamic state, accessibility semantics, deceptive microcopy [33], prompt-injection exposure [45], privacy-preserving personalization [84], and regulated decision support [74] were outside scope. Rubric design [85], and response-strategy control [86] may also change how a future critic communicates repairs. Calibration was empirical rather than conformal and may require refitting under distribution shift.

6. Conclusion

Uncertainty-aware selection improved Design2Code reconstruction by adapting generation strategy to each page. The selector surpassed all fixed policies, while its routing pattern showed that critique-repair is valuable when invoked selectively rather than universally. Complexity analysis and the concentration of failures in DOM fidelity identify structural reconstruction—not surface appearance—as the main remaining bottleneck.

Cross-fitted calibration converted unreliable raw confidence into an effective risk signal and improved selective performance. Together, rendering, DOM-aware evaluation, adaptive routing, and calibrated abstention provide a coherent workflow for static HTML/CSS generation. Extending this approach to contemporary models, interactive interfaces, responsive layouts, and shifted design distributions is the clearest path toward broader deployment.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The author wrote all of the sections in the article.

Ethics Statement

Ethical approval and permission were obtained from the Department of Foreign Languages and the relevant course lecturers by 25 July 2026

Informed Consent Statement

Written informed consent was obtained from all participants prior to their participation.

Data Availability Statement

The data are not publicly available due to confidentiality restrictions but be available from the corresponding author upon reasonable request and with permission from the data owner.

Declaration of Generative AI and AI-Assisted Technologies

Generative AI (ChatGPT, OpenAI) was used solely to assist with language editing and improving the clarity of the manuscript. The authors reviewed, edited, and verified all content and took full responsibility for the final version of the manuscript.

References

[1] C. Si, Y. Zhang, R. Li, Z. Yang, R. Liu, and D. Yang, "Design2Code: Benchmarking Multimodal Code Generation for Automated Front-End Engineering," *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3956-3974, 2025. doi: 10.18653/v1/2025.naacl-long.199.

[2] H. Assoudi, "Dataset Preparation," *Natural Language Processing on Oracle Cloud Infrastructure*, pp. 171-247, 2024. doi: 10.1007/979-8-8688-1073-2_5.

[3] Tony Beltramelli, "pix2code," pp. 1-6, 2018. doi: 10.1145/3220134.3220135.

[4] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?," *Advances in Neural Information Processing Systems 37*, pp. 87874-87907, 2024. doi: 10.52202/079017-2789.

[5] Y. Gui, Z. Li, Y. Wan, Y. Shi, H. Zhang, B. Chen, et al., "WebCode2M: A Real-World Dataset for Code Generation from Webpage Designs," *Proceedings of the ACM on Web Conference 2025*, pp. 1834-1845, 2025. doi: 10.1145/3696410.3714889.

[6] R. Li, Y. Zhang, and D. Yang, "Sketch2Code: Evaluating Vision-Language Models for Interactive Web Design Prototyping," *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3921-3955, 2025. doi: 10.18653/v1/2025.naacl-long.198.

[7] K. Li, L. H. U, and S. Shang, "App2Exa: Accelerating Exact kNN Search via Dynamic Cache-Guided Approximation," *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pp. 3045-3053, 2024. doi: 10.24963/ijcai.2024/339.

[8] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. 38th Int. Conf. Machine Learning, PMLR, vol. 139, 2021, pp. 8748–8763.

[9] R. Smith, "An Overview of the Tesseract OCR Engine," *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pp. 629-633, 2007. doi: 10.1109/icdar.2007.4376991.

[10] M. Pawlik, and N. Augsten, "Tree edit distance: Robust and memory-efficient," *Information Systems*, vol. 56, pp. 157-173, 2016. doi: 10.1016/j.is.2015.08.004.

[11] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Research & Application*, vol. 30, no. 1, pp. 21-30, 2004. doi: 10.1002/col.20070.

[12] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks," *Advances in Neural Information Processing Systems 36*, pp. 48573-48602, 2023. doi: 10.52202/075280-2109.

[13] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, et al., "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769-6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[14] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, et al., "Self-Refine: Iterative Refinement with Self-Feedback," *Advances in Neural Information Processing Systems 36*, pp.

46534-46594, 2023. doi: 10.52202/075280-2019.
[15] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems 36*, pp. 8634-8652, 2023. doi: 10.52202/075280-0377.
[16] Z. Gou et al., "CRITIC: Large language models can self-correct with tool-interactive critiquing," in Proc. Int. Conf. Learning Representations, Vienna, Austria, 2024.
[17] L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *Advances in Neural Information Processing Systems 36*, pp. 46595-46623, 2023. doi: 10.52202/075280-2020.
[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511-2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153.
[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning, PMLR, vol. 70, 2017, pp. 1321–1330.
[20] F. Pinto, H. Yang, S. N. Lim, P. Torr, and P. Dokania, "Using Mixup as a Regularizer Can Surprisingly Improve Accuracy & Out-Of-Distribution Robustness," *Advances in Neural Information Processing Systems 35*, pp. 14608-14622, 2022. doi: 10.52202/068431-1062.
[21] D. Prinster, A. Liu, and S. Saria, "JAWS: Auditing Predictive Uncertainty Under Covariate Shift," *Advances in Neural Information Processing Systems 35*, pp. 35907-35920, 2022. doi: 10.52202/068431-2602.
[22] S. Goren, I. Galil, and R. El-Yaniv, "Hierarchical Selective Classification," *Advances in Neural Information Processing Systems 37*, pp. 111047-111073, 2024. doi: 10.52202/079017-3526.
[23] J. T. Chien, "Deep Neural Network," *Source Separation and Machine Learning*, pp. 259-320, 2019. doi: 10.1016/b978-0-12-804566-4.00019-x.
[24] A. N. Angelopoulos, and S. Bates, "Conformal Prediction: A Gentle Introduction," *Foundations and Trends® in Machine Learning*, vol. 16, no. 4, pp. 494-591, 2023. doi: 10.1561/2200000101.
[25] S. Yun, H. Lin, R. Thushara, M. Bhat, Y. Wang, Z. Jiang, et al., "Web2Code: A Large-scale Webpage-to-Code Dataset and Evaluation Framework for Multimodal LLMs," *Advances in Neural Information Processing Systems 37*, pp. 112134-112157, 2024. doi: 10.52202/079017-3560.
[26] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in Proc. Int. Conf. Learning Representations, Kigali, Rwanda, 2023.
[27] Y. Chen, and M. Li, "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*, vol. 4, no. 2, pp. 364-384, 2025. doi: 10.51903/jtie.v4i2.490.
[28] Y. Li, S. Lu, and L. Zhao, "LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–215, May 2025, doi: 10.51903/ijgd.v3i1.3661.
[29] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," *International Journal of Graphic Design*, vol. 3, no. 1, pp. 216-240, 2025. doi: 10.51903/ijgd.v3i1.3713.
[30] J. Mu, Y. Lu, and E. Hwang, "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*, vol. 4, no. 1, pp. 192-208, 2026. doi: 10.51903/ijgd.v4i1.3702.
[31] H. Tu, S. Zhao, and A. Zhou, "Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 210–226, May 2025, doi: 10.51903/ijgd.v3i1.3714.

[32] J. Zhang, "Adaptive user interface design for volleyball learning apps: Empirical evidence from Google Play reviews and mobile screen analysis," Int. J. Graph. Des., vol. 3, no. 1, pp. 175–195, May 2025, doi: 10.51903/ijgd.v3i1.3618.

[33] H. Xu, Y. Chen, and A. Med, "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*, vol. 4, no. 3, pp. 590-612, 2025. doi: 10.51903/jtie.v4i3.491.

[34] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 447-463, 2025. doi: 10.51903/jtie.v4i2.522.

[35] H. Wang, Y. Ren, and X. Chang, "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*, vol. 4, no. 2, pp. 425-446, 2025. doi: 10.51903/jtie.v4i2.523.

[36] J. Jin, "LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397–414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[37] K. Zhang, Y. Chen, and A. Qian, "Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes," Int. J. Graph. Des., vol. 3, no. 2, p. 395, Oct. 2025, doi: 10.51903/ijgd.v3i2.3710.

[38] B. Zhang, Y. Ren, and J. Zou, "LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[39] Z. Li, S. Zhou, and Z. Zhou, "Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.

[40] Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415–436, Oct. 2025, doi: 10.51903/ijgd.v3i2.3616.

[41] T. Ye, X. Chang, and E. Zhong, "Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224," Int. J. Graph. Des., vol. 3, no. 2, pp. 365–380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3701.

[42] B. Zhou, H. Wang, and X. Chang, "Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365–380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3696.

[43] C. Li, B. Zhou, and K. Gao, "Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–394, Oct. 2025, doi: 10.51903/ijgd.v3i2.3709.

[44] Y. Zhang, and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces," *International Journal of Graphic Design*, vol. 3, no. 1, pp. 214-229, 2025. doi: 10.51903/ijgd.v3i1.3722.

[45] W. Su, H. Rao, and E. Ma, "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*, vol. 4, no. 1, pp. 186-191, 2026. doi: 10.51903/ijgd.v4i1.3699.

[46] J. Nie, G. Liu, and L. Zhao, "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*, vol. 4, no. 1, pp. 179-185, 2026. doi: 10.51903/ijgd.v4i1.3703.

[47] G. Liu, S. He, and H. Wong, "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*, vol. 3, no. 1, pp. 196-213, 2025. doi: 10.51903/ijgd.v3i1.3723.

[48] J. Zhang, "Early Warning, Grade Prediction, and Teacher-Facing LLM-Ready Explanations Toward an Open Volleyball Course: Reproducible Evidence from Four Public Education Datasets," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 20-44, 2026. doi: 10.51903/jtie.v5i2.525.

[49] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 215-238, 2026. doi: 10.51903/jtie.v4i1.485.

[50] J. Jin, "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*, vol. 4, no. 3, pp. 625-648, 2025. doi: 10.51903/jtie.v4i3.529.

[51] S. Lu, and T. Zou, "Uncertainty-Aware Medical Vision–Language Classification on a Lightweight MedMNIST-Compatible Biomedical Patch Benchmark," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 01-19, 2026. doi: 10.51903/jtie.v5i2.530.

[52] G. Mi, T. Ye, and D. Wood, "A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572–589, Aug. 2025, doi: 10.51903/jtie.v4i3.492.

[53] S. Zhao, J. Bai, and D. Roberson, "Multi-Horizon GPU Demand Forecasting with Workload Semantics and Operational Risk Curves: An Empirical Study on Alibaba Clusterdata GPU Trace," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 544-571, 2025. doi: 10.51903/jtie.v4i3.498.

[54] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 306-324, 2025. doi: 10.51903/jtie.v4i1.546.

[55] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "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*, vol. 4, no. 2, pp. 502-520, 2025. doi: 10.51903/jtie.v4i2.536.

[56] Q. Xin, "LiDAR–Camera Object-Level Fusion for Multi-Target Tracking Using JPDA and EKF: A Reproducible Empirical Study on a PandaSet-Parameterised Five-Sequence Dataset," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 54-76, 2026. doi: 10.51903/jtie.v5i1.486.

[57] Y. Chen, and H. Xu, "Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access," *International Journal of Graphic Design*, vol. 4, no. 1, pp. 141-164, 2026. doi: 10.51903/ijgd.v4i1.3552.

[58] Q. Xin, "Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," *J-INTECH*, vol. 14, no. 02, pp. 215-231, 2026. doi: 10.32664/j-intech.v14i02.2228.

[59] A. Y. Zhang, "Exact minimax optimality of spectral methods in phase synchronization and orthogonal group synchronization," *The Annals of Statistics*, vol. 52, no. 5, 2024. doi: 10.1214/24-aos2424.

[60] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC)*, vol. 8, no. 2, pp. 247, 2026. doi: 10.28989/avitec.v8i2.3974.

[61] A. Fanelli, and L. Moscardelli, "Approximately Stable Matching," *Frontiers in Artificial Intelligence and Applications*, 2025. doi: 10.3233/faia251234.

[62] W. Su, S. Chen, and C. Zhao, "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*, vol. 4, no. 3, pp. 649-662, 2025. doi: 10.51903/jtie.v4i3.543.

[63] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 520-533, 2025. doi: 10.51903/jtie.v4i2.535.

[64] W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 327-340, 2026. doi: 10.51903/jtie.v5i1.549.

[65] Y. Chen, S. Zhou, and E. Lin, "Accounting-Aware Evidence Retrieval for Institutional Due Diligence of Tokenized Trade Receivable RWA," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 649-663, 2025. doi: 10.51903/jtie.v4i3.542.

[66] S. Meng, J. Chen, and I. Zheng, "LLM-Inspired Offline Reranking for Financial Search: Query Rewriting, Hybrid Retrieval, and Listwise Relevance Ranking on FiQA," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 361-378, 2026. doi: 10.51903/jtie.v5i1.537.

[67] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms," 2025. doi: 10.48550/arxiv.2511.19481.

[68] Yifan Zhang, and Hailey Zhang, "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*, vol. 4, no. 2, pp. 464-486, 2025. doi: 10.51903/jtie.v4i2.547.

[69] Z. Li, Kai Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 75-90, 2026. doi: 10.51903/jtie.v5i2.541.

[70] S. Zhou, Y. Chen, and K. Lee, "Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 60-74, 2026. doi: 10.51903/jtie.v5i2.544.

[71] X. Sun, Ziliang Samuel Zhong, and Q. Wu, "Retrieval-Grounded HDFS Log Anomaly Detection and Deterministic Failure Narrative Generation," *Journal of Computational Systems and Applications*, vol. 3, no. 1, pp. 15-30, 2026. doi: 10.64229/j6d7fr94.

[72] X. Chang, Y. Lu, and Z. S. Zhong, "Review-Grounded Explainable Recommendation with Faithfulness Evaluation on Amazon Reviews," *JEECS (Journal of Electrical Engineering and Computer Sciences)*, vol. 11, no. 1, pp. 9-22, 2026. doi: 10.54732/jeecs.v11i1.2.

[73] Q. Xin, "Behavior retrieval plus response generation for interpretable conversational personalized recommendation," IJEEPSE, vol. 9, no. 2, pp. 120–136, Jul. 2026, doi: 10.31258/ijeepse.9.2.120-136.

[74] J. Li, and A. Zhou, "Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces," *Rule of Law Studies Journal*, vol. 2, no. 2, pp. 91-108, 2026. doi: 10.64780/rolsj.v2i2.225.

[75] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," *Emerging Information Science and Technology*, vol. 6, no. 2, pp. 125-146, 2025. doi: 10.18196/eist.v6i2.31232.

[76] T. Ye, J. Mu, and J. Hunter, "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*, vol. 5, no. 1, pp. 178-199, 2026. doi: 10.51903/jtie.v5i1.503.

[77] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model

Deployment," *Journal of Information Systems and Informatics*, vol. 7, no. 3, pp. 2182-2195, 2025. doi: 10.51519/journalisi.v7i3.1170.

[78] Siming Zhao, Y. Ren, and Xiaohan Chang, "Profit-Aware Spot GPU Admission Control with Cost-Sensitive Loss and Evidence-Grounded Policy Memos for AI Workload Supply-Demand Matching," *Journal of Technology Informatics and Engineering*, vol. 5, no. 2, pp. 45-59, 2026. doi: 10.51903/jtie.v5i2.545.

[79] S. He, J. Nie, and C. Li, "Power-Aware Inventory Planning for AI Infrastructure Using Job-Level Forecasting and LLM Workload Explanations," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 341-359, 2026. doi: 10.51903/jtie.v5i1.548.

[80] Z. S. Zhong, Chenyu Li, and H. Rao, "Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting on ZClawBench," *Journal of Technology Informatics and Engineering*, vol. 5, no. 1, pp. 341-360, 2026. doi: 10.51903/jtie.v5i1.539.

[81] C. Li, G. Liu, and Z. Zhao, "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*, vol. 5, no. 2, pp. 91-103, 2026. doi: 10.51903/jtie.v5i2.538.

[82] C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, "GPU Memory Requirement Prediction for Deep Learning Task Based on Bidirectional Gated Recurrent Unit Optimization Transformer," *2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV)*, pp. 31-35, 2025. doi: 10.1109/aivrv67401.2025.11350369.

[83] J. Bai, S. Chen, D. Zheng, and M.-J. Kuo, "Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing," Inf. Electr. Electron. Eng., vol. 6, no. 1, pp. 28–43, May 2026, doi: 10.33474/infotron.v6i1.24923.

[84] M. J. Kuo, D. Zheng, and J. Hires, "Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 385-401, 2025. doi: 10.51903/jtie.v4i2.502.

[85] Q. Xin, "Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education," J. Appl. Artif. Intell. Educ., vol. 2, no. 1, pp. 1–19, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[86] "Strategy-Aware Therapist Imitation for Emotional Support Dialogues: A Reproducible ESConv Study for LLM Response Control," *Advances in Educational Technology and Psychology*, vol. 10, no. 2, 2026. doi: 10.23977/aetp.2026.100213.

[87] Z. Samuel Zhong, and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," *Information and Inference: A Journal of the IMA*, vol. 13, no. 3, 2024. doi: 10.1093/imaiai/iaae020.

[88] Q. Xin, "Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction," Interdiscip. J. Pedagog. Res. Media Technol., vol. 2, no. 1, pp. 9–27, Jun. 2026, doi: 10.64268/inspire.v2i1.117.