WEBQX: Disclosing the Misalignment of Human Quality and Dense Retrieval in Webpage Optimization

Anonymous ACL submission

Abstract

Webpages increasingly serve two audiences: *humans*, who judge credibility and usefulness, and *machines*, which surface pages in retrieval-augmented generation (RAG) pipelines. Yet it remains unclear how improving a page for human readers affects its visibility to dense retrievers. To study this question, we introduce WEBQX, a three-part framework built on the *WebQuality* dataset of 60k webpages annotated along five human-centric dimensions. The framework contains: (1) WEBQX-Estimator, which predicts perceived quality from structural HTML features and exposes feature-level weaknesses using SHAP explanations; (2) WEBQX-OptAgent, a two-agent LLM pipeline that performs targeted HTML rewrites guided by these explanations; and (3) WEBQX-RAGEval, a retrievability evaluation module that evaluates how SHAP-guided HTML edits affect dense retrievability. Our experiments show that although SHAP-guided rewrites consistently improve predicted human quality, they systematically *degrade* dense retrieval performance at both page- and index-level metrics. Together, these results provide the first large-scale evidence of a structural misalignment between human-centered improvements and dense retrievability, highlighting the need for joint optimization strategies in RAG-mediated web access. We will release the code and trained components for reproducibility: https://anonymous.4open.science/r/webqxaisq-B38F/README.md.

1 Introduction

Retrieval-Augmented Generation (RAG) underpins much of today's web-facing language technology. Large Language Models (LLMs) increasingly ground their outputs in external webpages (Lewis et al., 2020; Gao et al., 2023), and modern systems expose source links that users frequently follow for verification. Webpage quality therefore shapes both user trust and retrieval effectiveness. A central question remains open: **How does human-perceived webpage quality relate to a page's retrievability under dense retrieval models?**

Prior work on *web quality assessment* has examined readability, credibility, structure, and content organization (Deepthi and Shalini, 2014; Castillo et al., 2011), while recent multimodal models (e.g., Hydra) produce learned quality scores (Zhang et al., 2025). In parallel, retrieval research has studied ranking robustness and grounding in RAG pipelines (Yu et al., 2024; Gao et al., 2023). Despite progress in both areas, their interaction has not been investigated yet.

To address this gap, we introduce **WEBQX** (see Figure 1), a three-part framework built on the WebQuality dataset (Zhang et al., 2025), which contains over 60k webpages annotated for human-perceived quality. First, we develop **WEBQX-Estimator**, a lightweight predictor of these dimensions that uses handcrafted structural features and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to expose which HTML properties drive low-quality assessments. We then use these explanations in **WEBQX-OptAgent**, a two-agent LLM pipeline in which a planner proposes targeted edits and an editor applies them to the HTML, following recent agentic optimization paradigms (Gong et al., 2025).

Finally, **WEBQX-RAGEval** evaluates how SHAP-guided HTML edits affect dense retrievability, comparing pre-/post-edit rankings at both page and index levels and validating trends on the multilingual *WebMMU* corpus (Awal et al., 2025).

In summary, our contributions are:

1. **WEBQX**: a unified, end-to-end framework that connects human-perceived webpage quality with dense retrievability through an interpretable feature space that combines handcrafted HTML features and retrieval-oriented

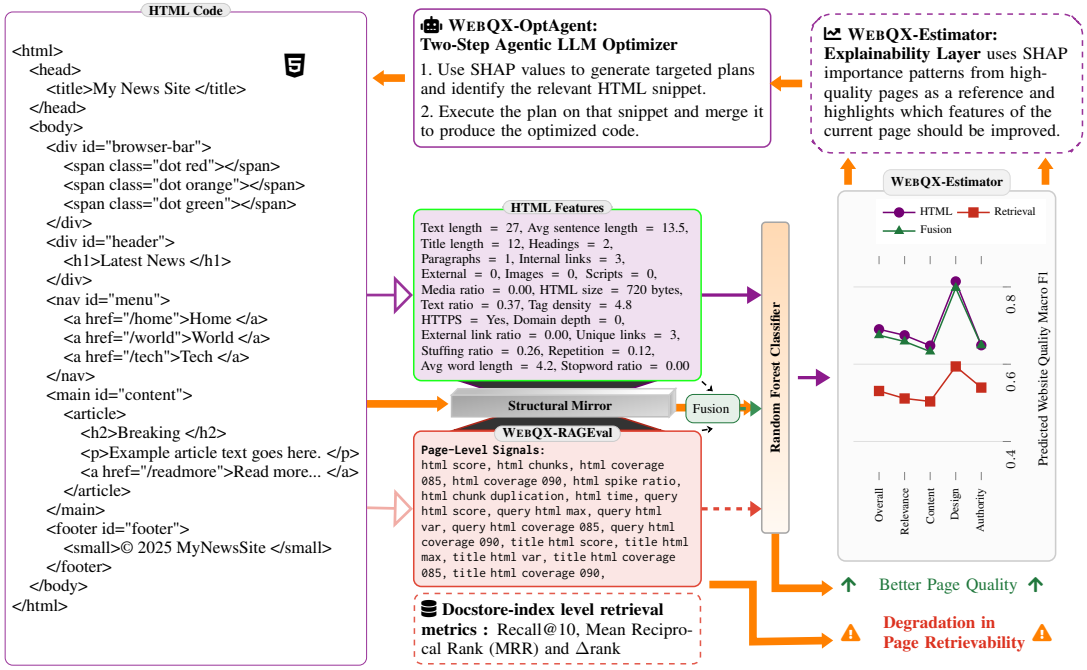


Figure 1: Full workflow of the proposed **WEBQX** framework. The original HTML page is transformed into structural and retrieval-based features. The HTML features capturing human-perceived quality are fed into **WEBQX-Estimator** to estimate quality scores, while retrieval-based features are used to evaluate retrievability in **WEBQX-RAGEval**. The **WEBQX-Estimator** also uses SHAP explanations to identify weaknesses, which are used in **WEBQX-OptAgent** to generate actionable rewrite plans and execute targeted edits, producing an optimized version whose quality and retrievability are evaluated against the baseline.

embedding metrics.

2. A SHAP-guided agentic optimization **WEBQX-OptAgent**: an integrated pipeline that leverages explanations from a lightweight, feature-based quality estimator to guide targeted HTML rewrites, consistently improving predicted human-perceived quality across five dimensions: *overall*, *relevance*, *content*, *design*, and *authenticity*.

3. Revealing a systematic trade-off in which human-centered structural optimizations that improve perceived quality consistently degrade dense retrieval performance in RAG systems.

Our findings reveal a consistent misalignment: although SHAP-guided rewrites improve predicted human quality, they systematically *harm* dense-retrieval performance. Rankings degrade, Recall@10 drops, and embedding similarity to the original query decreases. This suggests that dense retrievers rely on structural and lexical patterns disrupted by human-focused edits, highlighting the need for joint human–and–retriever optimization strategies in RAG-driven web ecosystems.

2 Related Work

Web Quality Assessment and Enhancement. Webpage quality has been studied across dimensions such as readability, credibility, design, and structure (Castillo et al., 2011; Deepthi and Shalini, 2014). Early approaches relied on handcrafted indicators (e.g., link patterns, keyword density, DOM structure, media usage) to approximate trustworthiness or professionalism (Hasan and Abuelrub, 2011; Morales-Vargas et al., 2023). More recently, large-scale annotated datasets such as *WebQuality* (Zhang et al., 2025) have enabled benchmarking multimodal quality prediction models (e.g., Hydra) that integrate textual, visual, and structural cues.

Prior work has also explored automated webpage enhancement, including design-oriented frameworks (Park and Noh, 2002), semantic enrichment for e-commerce (Necula et al., 2018), feedback-driven interaction improvements (Sufyan Beg, 2005), and transformer-based systems for rewriting or restructuring content (Vepsäläinen et al., 2024). In contrast to blind rewriting, our approach uses SHAP-based feature attributions to guide targeted and interpretable HTML edits.

Retrievability, Evaluation, and AI Visibility. RAG pipelines combine dense retrieval with large language models to improve factual grounding (Lewis et al., 2020; Gao et al., 2023). Retrieval performance is commonly evaluated using metrics such as recall@k, MRR, and nDCG (Yu et al., 2024; Alaofi et al., 2024), with recent work examining robustness, evidence sensitivity, and grounding fidelity (Li et al., 2024; Izacard et al., 2023; Upadhyay et al., 2024).

Related to this, Generative Engine Optimization (GEO) studies how content is adapted for visibility in AI-powered search systems (Aggarwal et al., 2024; Dai et al., 2025; Lüttgenau et al., 2025). Unlike GEO-focused work, our goal is not to optimize for generative search, but to quantify how human-centered quality improvements affect dense retriever accessibility.

Agents, Alignment, and Structural Awareness. Recent analyses show that LLM-based systems emphasize factual relevance over stylistic or credibility cues (Wan et al., 2024), suggesting divergences between human and machine notions of quality. Earlier work on web spam filtering (Cormack et al., 2011) similarly demonstrates that surface-level HTML structure can strongly influence retrieval behavior. Our work connects these insights by empirically measuring how HTML-level edits simultaneously affect human-perceived quality and dense retrievability in modern RAG settings.

3 Approach

First, we construct an interpretable feature space that links HTML structure, content presentation, and dense retrievability (§3.1). This space includes (i) HTML-derived features aligned with human-perceived quality (§3.1.1) and (ii) retrieval-oriented signals that capture how dense embeddings encode page similarity (§3.1.2).

Second, we introduce **WEBQX**, a three-component framework: (1) WEBQX-*Estimator*, which predicts human-perceived quality and provides SHAP explanations for each prediction (§3.2); (2) WEBQX-*OptAgent*, a two LLM agents (Planner and Executor) that uses these SHAP explanations to optimize the webpage for human-perceived quality (§3.3); and (3) WEBQX-*RAGEval*, which evaluates dense retrievability to assess how human quality-oriented edits affect retrieval performance (§3.4).

3.1 Feature Space

To create the feature space, we extract two main groups of indicators: (1) handcrafted HTML features designed to capture human-perceived webpage quality, and (2) retrieval-oriented metrics that approximate how dense retrievers score and rank pages. These two families capture complementary aspects of a webpage: its HTML organization and its retrieval-facing semantic footprint. We test whether dense retrievers exhibit **structural mirroring**, i.e., retrieval models do not operate on semantics alone, but may inherit biases from structural regularities in the underlying HTML. The full analysis appears in Section 5.1.

3.1.1 HTML Structure and Content Features.

Following prior work on web quality assessment (Deepthi and Shalini, 2014; Castillo et al., 2011; Hasan and Abuelrub, 2011; Morales-Vargas et al., 2023), we create an interpretable set of 21 HTML features capturing (i) *structural composition* (e.g., tag diversity, DOM depth, semantic balance), (ii) *content density and readability* (e.g., text–markup ratios, paragraph balance, media frequency), and (iii) *connectivity and authority* (e.g., internal/external link ratios, anchor richness, and domain metadata such as HTTPS and path depth). These features quantify how information is structured, presented, and interlinked—dimensions relevant to both human quality judgments and machine retrievability. An overview appears in Figure 1 (middle-middle), with full feature descriptions in Appendix A.1.

3.1.2 Retrieval-Driven Page-Level Signals.

To characterize retrieval-relevant signals, we compute embedding-based similarity features. Each webpage is segmented into textual chunks and paired with its title and associated query. To increase query coverage, we additionally generate five synthetic questions per page using `GPT-OSS-120B` (OpenAI, 2025), and compute embeddings with `moka-ai/m3e-base` (Wang Yuxin, 2023) (details in Appendix A.2). We derive features capturing *similarity* (query–chunk cosine similarity), *coverage* (fraction of chunks above similarity thresholds of 0.85 and 0.90), and *redundancy* (intra-page similarity and repetition patterns). These features serve as proxies for how easily dense retrievers can identify semantically relevant content. Their distributions are shown in Figure 1 (middle-bottom).

Example of LLM-generated recommendation on a random bad page

Feature: heading_count

Problem: 13 headings for only 268 words create a shallow structure and may be perceived as spammy.

Why it hurts overall quality: Too many headings fragment the conceptual structure, making it harder to extract a coherent summary or determine topical focus. It can also appear "spammy" to quality-assessment models.

Fixes:

- **Consolidate headings:** Merge closely related subsections under a single heading, keeping depth to 2–3 levels (H1 → H2 → optional H3).
- **Ensure logical hierarchy:** Each H2 should represent a distinct major topic; avoid using headings for short decorative sentences.
- **Use `aria-level`** attributes if visual styling is needed without altering the semantic structure (e.g., for subtitles).
- **Add structured data:** Include an outline in JSON-LD (`'@type': "WebPage"` with `breadcrumb` and `hasPart` entries) so agents can infer hierarchy even with simplified markup.

Figure 2: Planner Agent snippet illustrating our pipeline's SHAP-driven optimization recommendations.

3.2 Human-perceived Quality Prediction and Explanation: WEBQX-Estimator

To build the Estimator, we train Random Forest classifiers on the 21 handcrafted HTML features extracted from *WebQuality* dataset (Zhang et al., 2025) to predict scores for five Human-perceived Quality dimensions: *overall*, *relevance*, *content*, *design*, and *authenticity* (see Section 4.1). We train a separate classifier for each dimension.

To explain the predictions, we apply SHAP to each trained model to obtain per-sample, per-feature contribution scores across all output classes. We then aggregate these scores to produce both global feature-importance profiles and fine-grained, page-level attributions, highlighting which HTML features most strongly influence predicted human-perceived quality.

Using the Estimator, we obtain five predicted quality scores for each webpage, along with explanations for each predicted score.

3.3 Human-perceived Quality Optimization: WEBQX-OptAgent

To convert explanations into actionable improvements, we first transform the Estimator's raw SHAP explanations into a prioritized set of optimization targets. We then introduce **WEBQX-OptAgent**, a framework that leverages the estimator's predictions and SHAP explanations to improve human-perceived webpage quality. The framework comprises two LLM agents: a *Planner*, which translates SHAP feature attributions into structured optimization steps, and an *Executor*, which applies these edits at snippet- or page-level granularity.

3.3.1 Selecting Optimization Targets from SHAP Attributions

Given the raw SHAP feature attributions from the Estimator, we process each labelled and estimated low-quality webpage by ranking potential edit targets based on a combination of negative feature attributions and their corresponding raw feature values (e.g., an excessively high `heading count`). To determine the desired direction of change, we calculate the distribution of attribution differences between high- and low-quality pages across the dataset (see Appendix D.2).

This procedure outputs a prioritized shortlist of concrete, actionable structural issues, e.g. excessive headings, low external link density, or high content repetition, which serves as input to the *Planner* agent.

3.3.2 Optimization Agents

WEBQX-OptAgent consists of two coordinated LLM agents:

1. Planner Agent. Using SHAP tables, feature context, and attribution differences—following (Gong et al., 2025)—the Planner generates a structured JSON plan that specifies features, severity, rationale, fix type, and a precise description. An example output snippet appears in Figure 2, with further implementation details in Appendix E.1.

2. Executor Agent. For each planned action, the Executor locates the relevant HTML region, rewrites the snippet according to the specification, and outputs the modified fragment and a changelog. Fragments are then merged back into the full document, enabling controlled structural edits without breaking layout integrity.

3.4 Retrievability Assessment: WEBQX-RAGEval

Optimized pages are evaluated with **WEBQX-RAGEval**, which assesses their behavior in dense retrieval systems at two complementary levels. At the *page level*, we analyze retrieval-facing signals

WebQuality (Chinese)			WebMMU (Multilingual)		
HTML Feature	**Retrieval Feature**	r	**HTML Feature**	**Retrieval Feature**	r
repetition_score	title_html_max	-0.535	external_link_ratio	html_sim_max	$+0.405$
external_link_ratio	title_html_max	$+0.521$	external_link_ratio	title_html_max	$+0.387$
external_link_ratio	html_sim_max	$+0.473$	media_ratio	title_html_coverage_090	$+0.368$
stuffing_ratio	title_html_max	-0.463	title_len	title_html_max	$+0.360$
repetition_score	html_sim_max	-0.454	text_ratio	title_html_max	-0.328
external_link_ratio	query_html_max	$+0.405$	text_ratio	html_sim_max	-0.323
repetition_score	query_html_max	-0.374	external_link_ratio	html_sim_var	$+0.322$
stuffing_ratio	html_sim_max	-0.367	heading_count	html_sim_max	$+0.317$
stuffing_ratio	html_spike_ratio	-0.319	external_link_ratio	title_html_var	$+0.311$
tag_density	title_html_var	$+0.314$	external_link_ratio	html_spike_ratio	$+0.307$

Table 1: Top Pearson correlations between handcrafted HTML features and retrieval-derived features for two datasets: **WebQuality** (Left), **WebMMU** (Right). The latter supports four languages (English, Spanish, German, and French). Description of all features can be found in Appendix A

derived from dense embeddings to capture how human-centered HTML edits alter a page's representation. At the *document-store level*, we measure the downstream impact of these changes by inserting pages into a FAISS-based index (Douze et al., 5555) and selectively replacing embeddings of optimized low-quality pages while keeping the rest of the corpus fixed.

Retrievability Metrics. At the *page level*, **WEBQX-RAGEval** tracks retrieval-facing features rather than ranking outcomes. These include embedding drift, tokenization effects, and structural indicators such as link density and heading hierarchy. Shifts in these quantities reveal how local HTML modifications perturb the signals implicitly used by dense retrievers.

At the *index level*, retrievability is evaluated in a document-as-query setting. Each page's associated query is embedded and used to probe the index, ranking documents by inner-product similarity. We report rank position, Recall@10, and Mean Reciprocal Rank (MRR) before and after optimization. Recall@10 measures whether a page appears among the top-10 retrieved results for its query, while MRR captures overall ranking quality across pages. We additionally report Δrank, quantifying the per-page change in rank induced by optimization.

4 Experimental Setup

This section describes our experiments on predicting and optimizing human-perceived webpage quality and its effect on dense retrieval in RAG systems. We first introduce the WebQuality and WebMMU datasets (§4.1), then outline the *experimental setup* for analyzing correlations between HTML and retrieval features (§4.2), evaluating Human-perceived Quality Prediction (§4.3) and Optimization (§4.4), and assessing the impact of Human-Centered Optimization on RAG retrievability.

4.1 Datasets

Our primary dataset is *WebQuality* (Zhang et al., 2025), containing 60,000+ webpages annotated along five subjective dimensions: *overall*, *relevance*, *content*, *design*, and *authenticity*. Each dimension is rated on a three-level ordinal scale (**0 = poor**, **1 = ordinary**, **2 = excellent**), reflecting increasing levels of perceived quality. Each webpage includes its full HTML, extracted text, page title, and an associated user query, enabling analysis of both intrinsic quality and query-level relevance—an essential retrieval signal. To test cross-domain generalization, we additionally use the multilingual *WebMMU* corpus (Awal et al., 2025). Although unannotated for quality, it supports validation of structural correlations between quality-related features and retrievability.

4.2 Correlation Between HTML Features and Retrieval Signals

To examine how dense retrievers relate to webpage structure, we compute pairwise **Pearson correlations (r)** across all 378 combinations of retrieval-derived and handcrafted HTML features. Following common practice, we interpret $|r| > 0.3$ as moderate and $|r| > 0.5$ as strong associations.

4.3 Human-perceived Quality Prediction

We train the Random Forest classifiers of Estimator with z-score normalization, grid-searched hyperparameters, and 10-fold cross-validation (Appendix C). We report results as Macro-F1 ($\pm$ std).

Model	Overall Quality	Relevance	Professionalism	Design	Authenticity
MLP	0.62	0.60	0.59	0.72	0.56
Ordinal Model	0.62	0.62	0.60	0.77	0.56
Hydra (Zhang et al., 2025)	0.67	0.60	0.42	0.51	0.53
Hydra **+ H-HTML**	0.57	0.56	0.58	0.73	0.62
Hydra + H-HTML **w. mlp**	0.52	0.54	0.57	0.69	0.59
Hydra + H-HTML **w. attn**	0.52	0.52	0.53	0.56	0.59
WEBQX-Estimator (ours)	**0.69**	**0.68**	**0.65**	**0.81**	**0.65**

Table 2: Page-quality prediction results of **WEBQX-Estimator** across the five WebQuality dimensions. We compare standard ML baselines, the multimodal Hydra model (Zhang et al., 2025), and Hydra variants augmented with handcrafted HTML features via concatenation, MLP-based fusion, or attention-based fusion. Reported scores are Macro-F1 on the held-out test set; low cross-validation standard deviations on the training folds (0.006–0.018) further demonstrate stability (see Appendix C).

For baselines, we include MLPs (Popescu et al., 2009) as a strong tabular neural baseline, ordinal models (Winship and Mare, 1984) to model the ordered WebQuality labels (bad, ordinary, excellent), and the state-of-the-art multimodal model Hydra (Zhang et al., 2025) for webpage quality assessment. We extend Hydra by injecting the same handcrafted features via three fusion strategies: (i) projection into the shared embedding space, (ii) MLP-based concatenation before latent compression, and (iii) residual appending after multimodal fusion.

4.4 Human-perceived Quality Optimization

Following §3.3.1, we select all low-quality pages shorter than 100k characters to remain within the context limits of `GPT-OSS-120B` [1], yielding 1,100 pages for the optimization. These pages are re-scored by **WEBQX-Estimator**. We compute $\Delta Q = Q_{\text{after}} - Q_{\text{before}}$ (with Q for the quality of the web page) per dimension and summarize transitions in Table 3.

4.5 Impact of Human-Centered Optimization on RAG Retrievability

WEBQX-RAGEval evaluates the impact of human-centered edits on dense retrievability at two dimensions.

At the *page level*, it compares pre- and post-optimization retrieval features, such as embedding similarity, coverage, redundancy, and chunk structure, to quantify embedding-level shifts.

At the *index level*, we compute rank, Recall@10, MRR, and Δrank, and report the fraction of pages that improve, degrade, or remain unchanged. We use high-quality pages as an empirical upper bound for retrieval performance.

[1] https://platform.openai.com/docs/models/gpt-oss-120b

5 Results

This section presents the results of the experiments described earlier, including Correlation Between HTML and Retrieval indicators (§5.1), Quality Prediction (§5.2), Quality Optimization (§5.3), and the impact of Human-Centered Optimization on RAG Retrievability (§5.4), and highlights our key findings.

5.1 Dense Retrievers Systematically Reflect HTML Structure

Table 1 shows that retrieval similarity and coverage features—capturing query– and title-to-page alignment in embedding space (e.g., *title_html_max*, *html_sim_max*)—correlate strongly with structural HTML properties such as content repetition, external linking, and keyword density ($|r| \approx 0.45$–0.53). This indicates that dense retrievers systematically reward structurally explicit and redundant pages, rather than semantic relevance alone.

The same pattern holds in the multilingual *WebMMU* corpus (Awal et al., 2025), where retrieval signals correlate with indicators of hyperlink density, media usage, and text-to-markup balance (*external_link_ratio*, *media_ratio*, *text_ratio*; $|r| \approx 0.30$–0.40), indicating that these HTML-level structural biases generalize across languages.

5.2 HTML Structure Enables Accurate and Interpretable Quality Prediction

Table 2 shows that **WEBQX-Estimator** consistently outperforms all baselines across the five quality dimensions, achieving macro-F_1 scores ranging from 0.65 to 0.81, compared to 0.42–0.67

Target	Change Type	Overall Quality	Relevance	Professionalism	Design	Authenticity
Overall Quality (n=368)	0→1	**53.26**	49.46	19.84	8.42	6.25
	1→2	–	1.09	**10.87**	0.00	0.00
	0→2	**20.38**	2.27	4.35	0.00	0.00
	Unchanged	**26.36**	45.92	64.95	82.34	91.85
	↓	–	1.36	0.00	9.24	1.90
Relevance (n=147)	0→1	46.26	**71.43**	0.00	1.36	0.00
	1→2	0.00	–	**17.01**	0.00	0.00
	0→2	**18.37**	3.40	0.00	0.00	0.00
	Unchanged	35.37	**25.17**	82.99	93.20	99.32
	↓	0.00	–	0.00	5.44	0.68
Professionalism (n=33)	0→1	63.64	0.00	**78.79**	6.06	0.00
	1→2	0.00	**3.03**	–	0.00	0.00
	0→2	**24.24**	0.00	18.18	0.00	0.00
	Unchanged	12.12	93.94	**3.03**	78.79	100.00
	↓	0.00	3.03	–	15.15	0.00
Design (n=522)	0→1	13.76	1.92	2.11	**26.05**	1.72
	1→2	3.64	0.77	3.26	–	0.00
	0→2	**1.72**	0.00	0.00	0.00	0.00
	Unchanged	**74.33**	96.93	94.44	73.95	97.89
	↓	6.51	0.38	0.19	–	0.38
Authenticity (n=30)	0→1	36.67	0.00	0.00	20.00	**86.67**
	1→2	6.67	3.33	13.33	0.00	–
	0→2	**6.67**	0.00	0.00	0.00	0.00
	Unchanged	43.33	96.67	86.67	73.33	**13.33**
	↓	6.67	0.00	0.00	6.67	–

Table 3: Cross-dimensional quality transition matrix after targeted optimization. Each block corresponds to pages optimized along a given primary dimension (left). Rows indicate the type of change in categorical quality levels (**0→1**, **1→2**, **0→2** = improvements; **Unchanged**; ↓ = degradations), while columns represent the impact of that optimization on other quality dimensions. Values denote the percentage of affected pages among those initially rated as low quality (n = number of bad pages for that dimension).

for Hydra. These results indicate that models leveraging handcrafted HTML structural features yield more robust and stable predictions than architectures relying purely on multimodal or learned representations, confirming their effectiveness for modeling human-perceived webpage quality (Appendix A.3). Beyond predictive accuracy, **WEBQX-Estimator** provides explicit SHAP-based feature attributions that reveal how structural and textual cues—such as heading depth, link ratios, and repetition density—contribute to each quality dimension; these explanations directly support our optimization pipeline, where SHAP-informed signals guide LLM-based rewriting of low-quality pages. Additional attribution analyses and case studies are reported in Appendix D.

5.3 SHAP-Guided HTML Edits Reliably Improve Human-Perceived Quality

Table 4 shows that across 1,100 optimized pages, SHAP-guided edits consistently improve the targeted dimension, primarily through low-to-medium transitions (0→1) and fewer direct low-to-high jumps (0→2), with regressions being rare. For instance, optimizing `relevance` improves its own label on 74.8% of pages (71.43% via 0→1 and 3.4% via 0→2), while 25.2% remain unchanged. Cross-dimension effects are modest but positive (e.g., relevance score optimization improves the overall score on 64.63% of pages), whereas the page design remains the least responsive.

Ablation Settings. We measure the value of SHAP guidance, comparing against two reduced variants: (i) **No-SHAP** (features and references but no attributions), and (ii) **No-SHAP-No-Feature** (only a textual dimension description). As reported in 9, the SHAP-guided pipeline outperforms both baselines on four of five dimensions, showing that attribution signals materially support effective edits. The sole exception is *designScore*, where No-SHAP-No-Feature slightly leads, indicating that design cues are not well captured by HTML-level SHAP features (see Appendix E for the details).

5.4 Human-Centered Optimization Degrades Dense Retrievability

At the page level. Most retrieval-driven signals degrade following optimization. Figure 3 shows that signals related to content segmentation and structural regularity —such as chunk count, processing-time proxies, and spike ratios —de-

Metric	Overall Quality	Relevance	Professionalism	Design	Authenticity
Bad pages (#)	55,062	19,845	7,560	53,991	6,678
Good pages (#)	133,182	279,909	225,414	309,834	263,256
Good mean rank	3.07	3.24	3.04	3.12	3.25
Bad mean rank (orig)	2.97	2.88	2.70	3.53	3.02
Bad mean rank (opt)	3.25	3.32	2.99	3.77	3.02
Recall@10 (good)	0.789	0.767	0.789	0.800	0.776
Recall@10 (orig)	0.762	0.717	0.683	0.749	0.704
Recall@10 (opt)	0.741	0.714	0.675	0.723	0.704
MRR (good)	0.512	0.494	0.521	0.510	0.494
MRR (orig)	0.555	0.561	0.586	0.454	0.486
MRR (opt)	0.516	0.506	0.571	0.420	0.484
Δ rank (mean)	-0.33	-0.44	-0.39	-0.25	0.00
Improved (%)	6.1	5.5	5.1	8.6	5.4
Degraded (%)	15.0	17.7	16.5	17.2	4.5

Table 4: Retrievability evaluation before and after optimization across quality dimensions. Mean values are reported for bad pages unless otherwise stated. Δ rank denotes the mean rank change (negative = worse).

crease for 67–84% of pages, indicating reduced alignment with retriever-preferred layouts. Measures of query–document semantic alignment, including page-, title-, and fusion-level similarity scores, also decline for 65–73% of pages. Only a few signals remain stable or improve, such as high-threshold coverage (unchanged for 46%) and chunk duplication (improving for 49%), highlighting tension between human-centered edits and dense retrieval signals (Appendix F).

At the index level. Table 4 shows that optimized pages exhibit consistently lower retrievability: mean rank worsens (up to Δrank -0.45), recall@10 drops by 4–8 points, and degraded pages (5–18%) outnumber improved ones (5–8%). In contrast, high-quality pages remain stable (mean rank ≈ 3.1, recall@10 ≈ 0.79), suggesting that optimization primarily disrupts retriever-relevant cues while improving human-facing quality. These trends persist across additional embedding models (Appendix F.2).

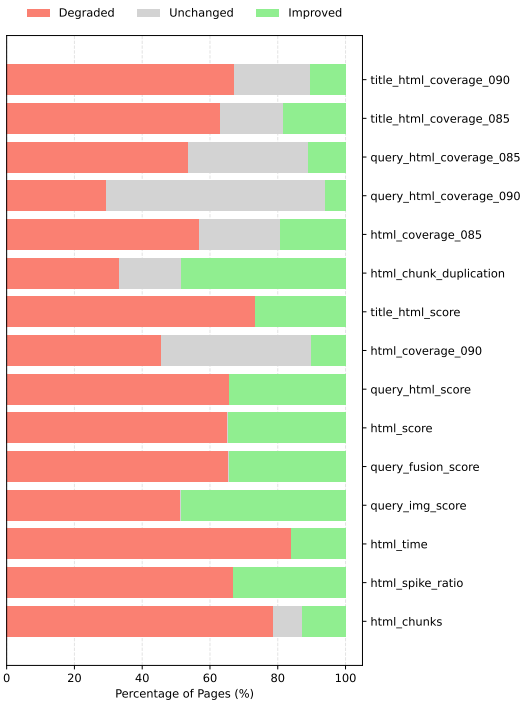


Figure 3: Overall Quality: Distribution of RAG feature shifts after SHAP-guided human-centered optimization. All structural and semantic retrieval features degrade for the majority of pages, indicating that edits that improve predicted human quality often reduce dense retriever compatibility. A few features, show minor improvements or remain stable, highlighting nuanced shifts in retrieval-relevant signals.

6 Conclusion

Taken together, our results provide the first empirical evidence that human-centered web optimization can unintentionally reduce visibility in retrieval-augmented systems. While optimized pages become objectively "better" for readers, they drift away from the structural and lexical regularities favored by dense retrievers, reducing both page-level and index-level retrievability. This work highlights a systematic misalignment between human-perceived quality and machine-centric relevance signals, suggesting that future web design and optimization frameworks must jointly consider readability, design, and retrievability. We hope these findings motivate new research on alignment-aware retrieval models and dual-objective web optimization techniques for the Web 4.0 ecosystem.

Limitations

Our work has several limitations that should guide interpretation. First, we operate on static HTML snapshots and do not model dynamic, interactive, or JavaScript-rendered content. Many modern webpages rely heavily on client-side rendering, and both human perception and retriever embeddings may shift once scripts execute or layouts adapt. Second, while Random Forests provide transparent SHAP attributions, they may miss subtle semantic or multimodal cues captured by transformer-based architectures. SHAP values reflect feature importance *within the RF*, and thus inherited model biases may shape the optimization plans. Third, the LLM executor is subject to context-length constraints. We therefore limit optimization to bad pages below a given character threshold (100k), which introduces a selection bias toward shorter or simpler pages. After rewriting, the quality of these pages is re-assessed using the same RF model—meaning that our "ground truth" improvement signal is an estimate and may favor features the RF already prefers. Fourth, we evaluate only the HTML markup and cannot verify the visual rendering or usability of the rewritten pages; an updated DOM may pass the RF checks yet produce degraded layout or accessibility issues. Fifth, we do not conduct a direct human evaluation of (pre- and) post-optimization pages. Running such a study at scale is challenging, especially given the predominance of Chinese-language content, but would provide a stronger validation signal than model-based estimates. Sixth, while we evaluate retrievability and demonstrate the misalignment between human-centered improvements and embedding-based retrieval, we do not propose a method to jointly optimize pages for both human and model preferences. Designing approaches that balance these objectives remains an important direction for future work. Finally, our primary dataset, *WebQuality*, is predominantly Chinese. While auxiliary tests on the multilingual *WebMMU* corpus suggest that several structural trends generalize, broader multilingual and domain-diverse evaluation is needed to establish external validity.

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. GEO: generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 5–16. ACM.

Marwah Alaofi, Negar Arabzadeh, Charles L. A. Clarke, and Mark Sanderson. 2024. *Generative Information Retrieval Evaluation*, page 135–159. Springer Nature Switzerland.

Rabiul Awal, Mahsa Massoud, Zichao Li, Aarash Feizi, Suyuchen Wang, Christopher Pal, Aishwarya Agrawal, David Vazquez, Siva Reddy, Juan A. Rodriguez, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. 2025. WebMMU: A benchmark for multimodal multilingual website understanding and code generation. In *ICLR 2025 Third Workshop on Deep Learning for Code*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 675–684. ACM.

Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Inf. Retr.*, 14(5):441–465.

Sunhao Dai, Wenjie Wang, Liang Pang, Jun Xu, See-Kiong Ng, Ji-Rong Wen, and Tat-Seng Chua. 2025. Next-search: Rebuilding user feedback ecosystem for generative ai search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3922–3931. ACM.

M. Deepthi and S. Shalini. 2014. Quantitative evaluation of web page attributes for improving the quality of website. *International Journal of Engineering Research & Technology (IJERT), NCRTS – 2014*, 2(13).

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazare, Maria Lomeli, Lucas Hosseini, and Herve Jegou. 5555. THE FAISS LIBRARY . *IEEE Transactions on Big Data*, (01):1–17.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Jingzhi Gong, Rafail Giavrimis, Paul Brookes, Vardan Voskanyan, Fan Wu, Mari Ashiga, Matthew Truscott, Mike Basios, Leslie Kanthan, Jie Xu, and Zheng Wang. 2025. Tuning llm-based code optimization via meta-prompting: An industrial perspective. *CoRR*, abs/2508.01443.

Layla Hasan and Emad Abuelrub. 2011. Assessing the quality of web sites. *Applied Computing and Informatics*, 9.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. 2024. Understanding retrieval robustness for retrieval-augmented image captioning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9285–9299. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Florian Lüttgenau, Imar Colic, and Gervasio Ramirez. 2025. Beyond SEO: A transformer-based approach for reinventing web content optimisation. *CoRR*, abs/2507.03169.

Alejandro Morales-Vargas, Rafael Pedraza-Jimenez, and Lluís Codina. 2023. Website quality evaluation: a model for developing comprehensive assessment instruments based on key quality factors. *Journal of Documentation*, 79(7):95–114.

Sabina-Cristiana Necula, Vasile-Daniel Păvăloaia, Cătălin Strîmbei, and Octavian Dospinescu. 2018. Enhancement of e-commerce websites with semantic web technologies. *Sustainability*, 10(6).

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925.

Hee-Sok Park and Seung J. Noh. 2002. Enhancement of web design quality through the qfd approach. *Total Quality Management*, 13(3):393–401.

Marius-Constantin Popescu, Valentina Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. 2009. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8.

M.M. Sufyan Beg. 2005. User feedback based enhancement in web search quality. *Information Sciences*, 170(2):153–172.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2024. A large-scale study of relevance assessments with large language models: An initial look. *CoRR*, abs/2411.08275.

Juho Vepsäläinen, Arto Hellas, and Petri Vuorimaa. 2024. Overview of web application performance optimization techniques. *CoRR*, abs/2412.07892.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7468–7484. Association for Computational Linguistics.

He sicheng Wang Yuxin, Sun Qingxuan. 2023. M3e: Moka massive mixed embedding model.

Chris Winship and Robert Mare. 1984. Regression models with ordinal variables. *American Sociological Review*, 49:512.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *CoRR*, abs/2405.07437.

Tao Zhang, Yige Wang, ZhuHangyu ZhuHangyu, Li Xin, Chen Xiang, Tian Hua Zhou, and Jin Ma. 2025. Webquality: A large-scale multi-modal web page quality assessment dataset with multiple scoring dimensions. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 583–596.

A Feature Specification

This appendix provides the full definition of all features used in our analysis. We include (i) the handcrafted HTML quality indicators used by the predictive RF model and by the SHAP-based optimization procedure, and (ii) the retrieval-driven features used to quantify how easily a dense retriever can extract relevant information from a webpage.

A.1 Handcrafted HTML Quality Features

Table 6 lists the handcrafted features grouped into five categories: textual content, structural layout, link and media structure, authenticity indicators, and authority/metadata signals. These features reflect properties that are commonly associated with human judgments of readability, coherence, and credibility. They are also used directly in the optimization pipeline, and the exact computation formulas are made available in the released codebase.

Generate {n} short user questions in Chinese about the content of this webpage.
Strict requirements:

1. All questions must be written in Chinese.
2. The output must be in valid JSON format: `{"questions": ["q1", "q2", ...]}`.

Webpage content:

`[CONTENT]`

Figure 4: Prompt used for `GPT-OSS-120B` generation

A.2 RAG-Derived Retrieval Features

To measure how HTML structure aligns with retrieval utility, we compute a set of embedding-based similarity metrics capturing how well each webpage's content can be matched to different forms of queries.

Chunking. Each webpage is segmented into a collection of textual units (paragraphs, headings, list items, and the page title). These chunks serve as candidates for dense retrieval.

Query Sources. Each page is associated with three query types: (i) the human-written query provided by the *WebQuality* dataset, (ii) the page title, and (iii) five automatically generated user questions. Synthetic questions are produced with `GPT-OSS-120B` using a simple language-specific prompt requesting short information-seeking questions (see Figure 4). For multilingual datasets (e.g., *WebMMU*), questions are generated in the corresponding page language using a model with multilingual coverage.

Embedding Models. We compute text embeddings using `moka-ai/m3e-base` for chunk-, query-, and title-level similarity. For screenshots, CLIP embeddings (`CLIP-ViT-B-32`) are used to measure cross-modal alignment.

Feature Categories. Retrieval-derived features are grouped into:

- **Similarity statistics:** mean, max, and variance of cosine similarity between queries and chunks.
- **Coverage:** proportion of chunks exceeding similarity thresholds (0.85, 0.90).
- **Redundancy:** intra-page chunk duplication and spike ratios.
- **Cross-modal relevance:** CLIP-based query–image and title–image alignment.
- **Fusion metrics:** combined HTML + screenshot similarity.

A complete specification of these RAG features is provided in Table 7.

A.3 Feature-Set Configurations Used in the Prediction Ablation Study

To determine the final design of **WEBQX-Estimator**, we evaluate seven feature configurations that differ in which modalities and retrieval signals are available to the model. Each configuration corresponds to a distinct subset of the page-level features computed by our RAG-based analysis pipeline (Section 4.1). Below we summarize what each configuration captures.

HTML-Only. This setting uses exclusively handcrafted structural and linguistic indicators extracted directly from the HTML source (e.g., text length, heading and paragraph counts, link structure, media density, repetition and stuffing ratios, domain-depth, HTTPS flag). These features describe a page's surface form, organization, and content redundancy without relying on embedding models.

RAG-Only. This configuration uses retrieval-derived signals obtained from embedding-based comparisons between (i) generated questions, (ii) the user query, (iii) the page title, and (iv) HTML content chunks. Features include cosine-similarity statistics (`html_sim_mean`, `html_sim_max`, `query_html_score`, `title_html_score`), coverage thresholds (0.85/0.90), chunk-duplication scores, spike ratios, and retrieval-time measurements. These signals reflect how the page aligns semantically with the generated diagnostic queries and with the ground-truth query/title.

Image-Only. This modality incorporates CLIP-based comparisons between the page screenshot and the generated questions, user query, and title. Features include similarity statistics (`img_score`, `img_sim_mean`) and per-query alignment scores (`query_img_score`, `title_img_score`), capturing visual–textual consistency.

HTML + RAG. A union of handcrafted HTML indicators and retrieval-based textual signals. This setting tests whether semantic alignment information improves predictions beyond structural cues.

HTML + Image. Combines HTML structural features with screenshot–text alignment features. This assesses whether visual–textual match contributes useful complementary information to structural signals.

RAG + Image. Uses only embedding-based textual and visual retrieval features. This configuration removes all handcrafted HTML indicators to evaluate whether learned representations alone can approximate human quality judgments.

All Features. A full multimodal fusion of HTML structure, textual retrieval signals, and screenshot features. This represents the richest feature set, testing whether combining all signals yields further gains or introduces noise.

Across configurations (Table 5), HTML-only consistently provides the strongest and most stable performance. Adding RAG or image signals does not improve prediction accuracy and in several dimensions degrades it, indicating that structural HTML cues remain the most informative and least noisy predictors of human-perceived quality. These findings motivate the design choice of **WEBQX-Estimator** as an HTML-only model.

B Full Correlation Matrices

For completeness, we provide the full Pearson correlation matrices covering all handcrafted HTML features and all retrieval-derived RAG features. While the main paper discusses only the strongest effects (Figures 1), the full matrices confirm the same structural patterns across both datasets.

WebQuality. Figure 5 reports the complete HTML–retrieval correlation matrix for the WebQuality corpus. The structure, verbosity, and link-related features show consistent correlations with retrieval similarity, coverage, and redundancy metrics. These global trends mirror the compact analysis in the main text, but also reveal smaller secondary dependencies that are not visible in high-level summaries.

WebMMU. Figure 6 provides the corresponding matrix for the multilingual WebMMU dataset. Despite differences in language and content diversity, the overall correlation landscape remains highly consistent with WebQuality. In particular, features related to chunk length, heading density, and internal redundancy show stable effects across languages, supporting the claim that HTML structure influences dense-retrieval behavior in a language-agnostic manner.

C Cross-Validation Results on the Training Split

Table 8 reports 10-fold cross-validation accuracy and macro-F_1 (mean ± std) for the HTML-only Random Forest models trained on the WebQuality development split. Overall, the scores indicate stable training dynamics, with low variance across folds. Design and Authenticity are predicted most reliably (acc. ≥ 0.88), suggesting that structural and aesthetic webpage cues are highly correlated with human judgment. In contrast, Overall Quality and Professionalism exhibit lower F_1 despite moderate accuracy, indicating more subjectivity and label imbalance that complicate prediction. These results reflect performance on the training split via 10-fold cross-validation and are not used for final evaluation; held-out test benchmarking, including comparison to the Hydra baseline, is reported in Section 4.3.

Table 8: Training-set 10-fold cross-validation performance (mean ± std).

Dimension	Acc. (CV)	F_1 (CV)
Overall Quality	0.701 ± 0.008	0.685 ± 0.008
Relevance	0.802 ± 0.007	0.685 ± 0.009
Professionalism	0.791 ± 0.007	0.651 ± 0.009
Design	0.882 ± 0.004	0.804 ± 0.006
Authenticity	0.920 ± 0.004	0.657 ± 0.018

D Extended SHAP Interpretability Analysis

This appendix provides the full explainability artefacts used throughout the Two-Step Agentic Optimization pipeline described in section 5.2. All SHAP computations are derived from the RF models trained for the five WebQuality dimensions: *Overall*, *Relevance*, *Content*, *Design*, and *Authenticity*. For each dimension, the classifier predicts a discrete quality score ($0, 1, 2$ for most labels; $0, 1$ for *Authenticity*), and SHAP values quantify feature-level contributions to those predictions.

D.1 Class-wise contribution structure

For each quality dimension, we visualise how individual features influence the probability of a page being assigned to each quality level (Figures 7–11). Features that appear near the top have the largest overall impact; colour encodes whether high or low

Table 5: F1 scores (rounded) for different feature combinations across quality dimensions.

Features	overScore	releScore	contScore	designScore	authScore
html_only	0.69	0.68	0.65	0.81	0.65
rag_only	0.53	0.51	0.50	0.59	0.54
img_only	0.36	0.37	0.36	0.37	0.51
html_rag	0.68	0.66	0.64	0.80	0.66
html_img	0.69	0.67	0.65	0.81	0.65
rag_img	0.54	0.52	0.51	0.61	0.54
all	0.68	0.66	0.63	0.80	0.65

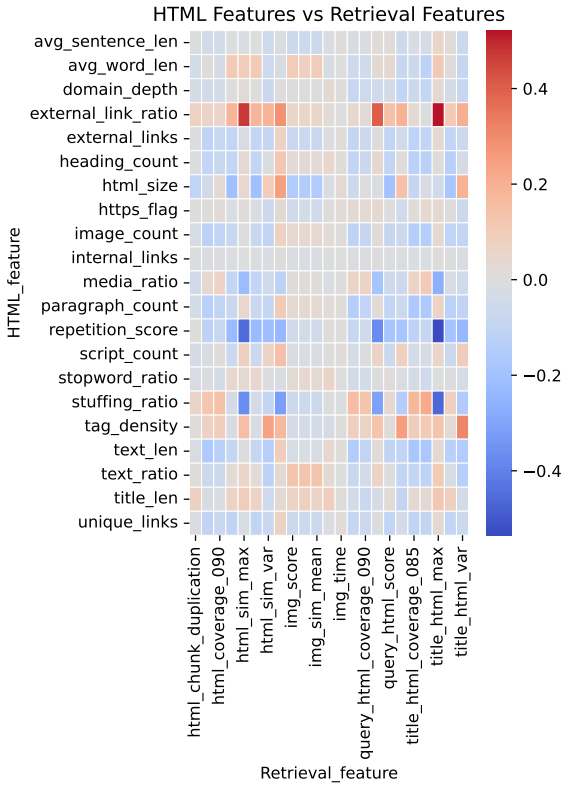


Figure 5: Full Pearson correlation matrix between HTML features and retrieval features in the WebQuality dataset.

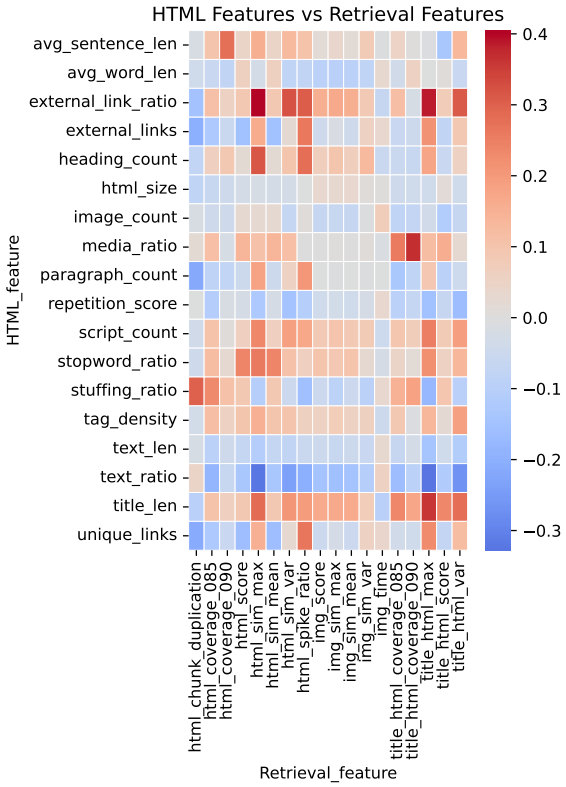


Figure 6: Full Pearson correlation matrix between HTML features and retrieval features in the WebMMU dataset.

feature values push predictions upward or downward. These plots reveal:

- which HTML characteristics consistently push documents toward the *low-quality* class;
- which structural cues are characteristic of *high-quality* pages;
- which signals differ across dimensions (e.g., strong for content, weaker for design).

D.2 Cross-page contrasts between good and bad pages.

To complement these global summaries, Figures 12–16 provide two additional per-dimension diagnostics:

- **ΔSHAP violin plots** (left panels), showing the distribution of attribution differences between good and bad pages;
- **ΔFeature-value bar plots** (right panels), showing mean differences in raw feature values (good − bad).

Table 6: Handcrafted HTML quality features used in our predictive model and SHAP-guided optimization pipeline. Colors denote feature groups.

Feature	Description	Why it Matters
Text Features		
text_len	Number of words extracted from visible text.	Longer text increases semantic coverage but excess verbosity may reduce clarity.
avg_sentence_len	Average number of words per sentence.	Signals readability; extreme values correlate with perceived low quality.
title_len	Character length of the HTML title.	Short or vague titles reduce relevance and weaken retriever alignment.
Structural Layout		
heading_count	Number of H1–H6 tags.	Indicates structural hierarchy; too many headings fragment content and look spammy.
paragraph_count	Count of <p> elements.	Reflects textual organization; sparse paragraphing harms readability.
Links and Media		
internal_links	Links pointing to same domain.	Supports navigation and authority signals; improves structural cohesion.
external_links	Links to outside domains.	Excess external linking can appear spammy and degrades retriever trust patterns.
image_count	Number of images.	Enhances design quality but too many images dilute textual similarity for retrievers.
script_count	Script tags embedded in page.	Heavy scripting reduces readability and harms reliability signals.
media_ratio	Media tags relative to all tags.	High ratios reduce content density and weaken semantic retrievability.
Complexity and Density		
html_size	Raw HTML length in bytes.	Large files reduce clarity; retrievers often favor compact, semantically dense pages.
text_ratio	Visible-text-to-HTML size ratio.	High ratios indicate cleaner pages; low ratios signal clutter and template noise.
tag_density	Total tags per 100 words.	Measures structural fragmentation; high density correlates with low-quality templates.
Authenticity and Redundancy		
stuffing_ratio	Fraction of text dominated by top-10 words.	Detects keyword stuffing, a strong indicator of low-quality or manipulative content.
repetition_score	Proportion of repeated trigrams.	Captures boilerplate or duplicated phrasing; affects both quality and retriever behavior.
avg_word_len	Average word length.	Long words raise sophistication; overly short tokens suggest templated or noisy content.
stopword_ratio	Frequency of stopwords (CN baseline set).	Extremely low ratios signal unnatural text or keyword stuffing.
Authority and Metadata		
https_flag	Whether page uses HTTPS.	HTTPS is a trust and security signal for humans and ranking systems.
domain_depth	URL path depth.	Deep paths often indicate outdated CMS structures or low-authority pages.
external_link_ratio	Fraction of links pointing externally.	High values signal spam or low authority; retrievers penalize such patterns.
unique_links	Number of distinct linked URLs.	Indicates page richness; very low uniqueness suggests templated duplication.

These views jointly express how structural and textual signals vary between quality levels. Across all dimensions, positive ΔSHAP features—including *paragraph_count*, *heading_count*, *text_ratio*, and *image_count*—indicate that high-quality pages tend to have richer structure, cleaner markup, and well-integrated media. Conversely, strongly negative features such as *stuffing_ratio*, *repetition_score*, *media_ratio*, and excessive *unique_links* mark low-quality pages and align with canonical failure patterns (keyword stuffing, redundant text blocks, link spam).

Dimension-specific differences further refine this picture. For *relevance* and *content*, natural

Table 7: RAG-derived retrieval features used in our analysis. Features are grouped by retrieval source: HTML chunks (generated-question alignment), real user queries, page titles, screenshots, and fusion metrics. Each feature includes its definition and relevance to embedding-based retrievability.

Feature	Description	Why It Matters
HTML–Retrieval Features (Generated Questions → HTML Chunks)		
`html_chunks`	Number of extracted text chunks (title, headings, paragraphs, list items).	Determines granularity of retrievable units; more chunks give the retriever more entry points.
`html_sim_mean`	Mean cosine similarity between generated questions and all chunks.	Measures overall semantic alignment of page content with information-seeking queries.
`html_sim_max`	Maximum similarity across all question–chunk pairs.	Indicates whether the page contains at least one highly relevant segment.
`html_sim_var`	Variance of similarities.	High variance suggests heterogeneous or inconsistent topical focus.
`html_coverage_085/09(`	Fraction of chunks above similarity thresholds (0.85 / 0.90).	Acts as a proxy for recall at high-precision cutoffs.
`html_spike_ratio`	Ratio of max similarity to mean similarity.	Detects "one good chunk" pages with otherwise weak content.
`html_chunk_duplicati(`	Percentage of chunk pairs with similarity > 0.9.	Captures redundancy and near-duplicate sections.
`html_time`	Time to compute all question–chunk similarities.	Reflects structural complexity and embedding overhead.
Query–Retrieval Features (Real Query → HTML Chunks)		
`query_html_score`	Mean similarity between the real query and chunks.	Measures semantic match for the actual user intent.
`query_html_max`	Max similarity for query–chunk pairs.	Detects whether the page contains a directly query-relevant span.
`query_html_var`	Variance of query–chunk similarities.	High variance indicates misaligned or unevenly distributed relevant information.
`query_coverage_085/0(`	Chunk coverage above similarity thresholds.	Useful for predicting recall within dense-retrieval pipelines.
`query_img_score`	CLIP similarity between the query and screenshot.	Captures query-to-visual relevance.
`query_fusion_score`	Fusion of HTML and image similarity for the query.	Represents full-page retrievability across modalities.
Title–Retrieval Features (Page Title → HTML Chunks)		
`title_html_score`	Mean similarity between title and chunks.	Indicates whether the title accurately summarizes page content.
`title_html_max`	Max similarity for title–chunk pairs.	Detects presence of a canonical, title-aligned main section.
`title_html_var`	Variance of title–chunk similarity.	High variance suggests mismatch between title and body content.
`title_coverage_085/0(`	Fraction of chunks strongly aligned with the title.	Important for title-based indexing and ranking heuristics.
`title_img_score`	CLIP similarity between title and screenshot.	Measures whether the visual appearance matches the stated topic.
`title_fusion_score`	Average of HTML and image relevance for the title.	Proxy for title consistency across modalities.
Image–Retrieval Features (Generated Questions / Query / Title → Screenshot)		
`img_score`	Mean CLIP similarity between generated questions and screenshot.	Measures visual informativeness with respect to content queries.
`img_sim_mean/max/var`	Summary statistics of question–image similarity.	Capture visual relevance and noise introduced by decorative elements.
`img_time`	Embedding time for the screenshot.	Indicates computational cost for multimodal RAG pipelines.
Fusion Features (Cross-Modal HTML + Image Aggregation)		
`fusion_score`	Combined HTML and image similarity for generated-question retrieval.	Represents overall retrievability when both modalities are used.
`query_fusion_score`	Fusion of query-based HTML and image alignment.	Approximates real-query ranking behavior under multimodal retrievers.
`title_fusion_score`	Fusion of title-based relevance across HTML and image modalities.	Useful for assessing title–page consistency.

text cues (*stopword_ratio*, *avg_word_len*) sharply distinguish good pages, whereas filler-heavy pages exhibit large negative SHAP values for *text_len* and *stuffing_ratio*. For *design*, visually coherent pages (higher *image_count*, balanced *text_ratio*) receive positive attributions, whereas media-heavy clutter or misused headings drive predictions downward. For *authenticity*, higher-quality pages exhibit stronger structural regularity (positive *tag_density*, *html_size*), while low-quality pages often feature fragmented layout or overly complex wording.

(3) Role in the optimization pipeline. Together, these SHAP views clarify which structural weaknesses most strongly characterize low-quality pages. This appendix therefore provides the supplementary interpretability material underlying our SHAP-guided optimization framework: the Planner Agent uses precisely these feature-level attributions and cross-page differences to generate targeted, feature-specific improvement actions for each page.

E Optimization Pipeline and Ablation Details

Building on the feature-level insights provided by the SHAP analyses in Appendix D.2, we construct a two-step agentic pipeline for human-centered webpage optimisation (Section 3.3). The pipeline consists of:

1. a **Planner Agent** that transforms SHAP explanations and feature statistics into a structured improvement plan, and
2. an **Editor Agent** that applies these actions to the HTML.

E.1 Planner Agent

The Planner Agent forms the first stage of the two-step optimisation pipeline. Its role is to translate interpretability signals into a structured set of actions that describe *what* should be improved in a low-quality page and *why*.

Prompt Construction. As described in Section 3.3, the agent receives three sources of contextual information:

1. **Feature explanations**: concise textual definitions of each HTML feature.
2. **Page-specific diagnostics**: the page's raw feature values and their SHAP contributions, indicating which attributes most influenced the model's low-quality prediction.
3. **Reference differences**: aggregated good–bad contrasts (ΔSHAP and Δvalue) from Appendix D.2, identifying globally important weaknesses.

To help the agent decide *where* in the HTML the proposed fixes should be applied, we additionally provide a small mapping from each feature to the HTML tags most relevant to it (e.g., `heading_count` → `<h1>-<h6>`, `text_ratio` → `<p>`, and `image_count` → `<img>`). These hints are generated automatically by a lightweight LLM prompt and act only as soft anchors, guiding the agent toward the appropriate code regions without prescribing specific edits.

Given the full context described above, the Planner Agent outputs a JSON list of optimisation actions that serve as the blueprint for the subsequent code-editing stage. The complete prompt is shown in Figure 17.

E.2 Editor Agent

E.3 Ablation Conditions

To quantify how much the Planner Agent benefits from SHAP-derived guidance, we evaluate two degraded versions of the prompt:

- **No-SHAP**: feature values and reference differences are provided, but all SHAP contributions are removed.
- **No-SHAP-No-Feature**: the agent receives only a textual description of the target quality dimension (e.g., "improve design quality"), with no structured features or numeric signals.

These settings isolate the contribution of the SHAP explanations and reveal how they improve the specificity and relevance of the generated optimisation plans. Table 9 results to be updated once the experiement are done

Example Planner Output. Below is a shortened example of the JSON produced for a low-quality page under the full (SHAP-enabled) configuration:

This example illustrates how the agent combines page-level diagnostics with global structural patterns to produce targeted, interpretable, and hierarchically organised optimisation instructions.

```
[
  {
```

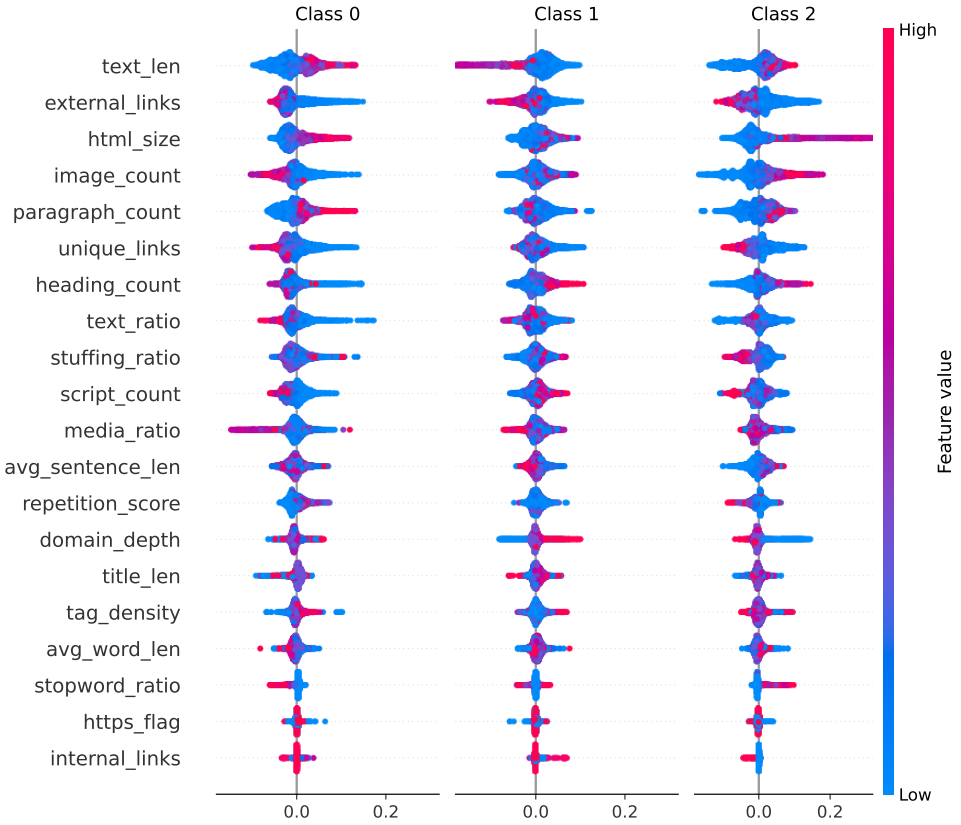


Figure 7: SHAP summary plots for the **Overall Quality** across all classes (0, 1, 2).

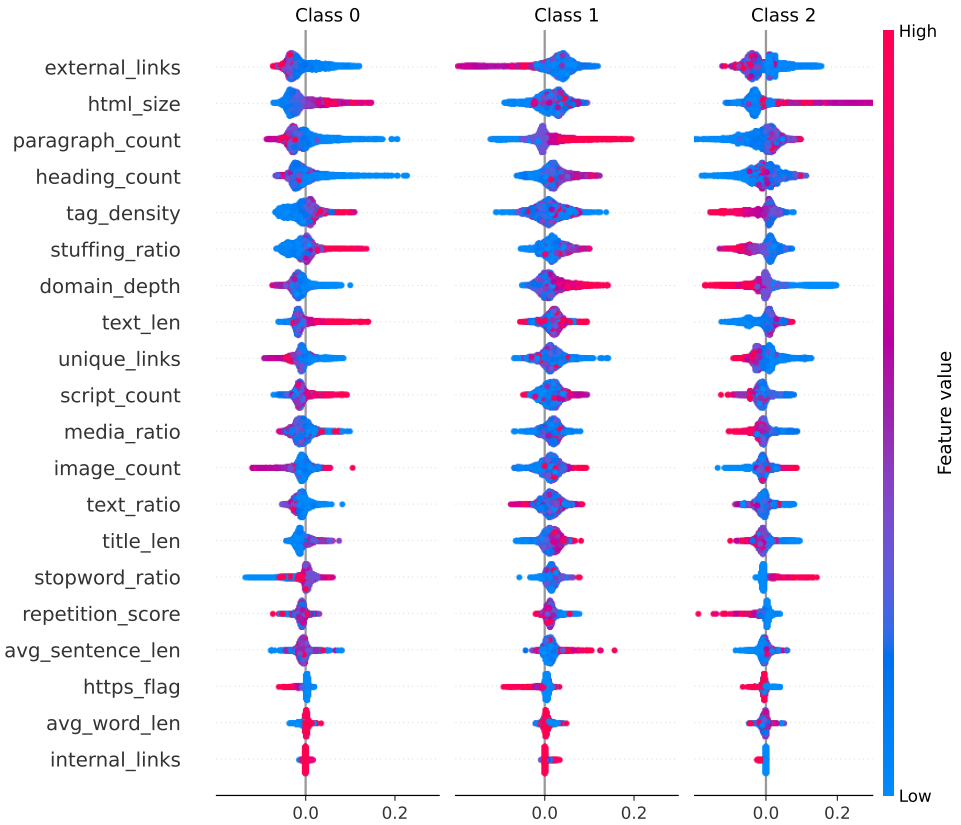


Figure 8: SHAP summary plots for the **Relevance** across all classes (0, 1, 2).

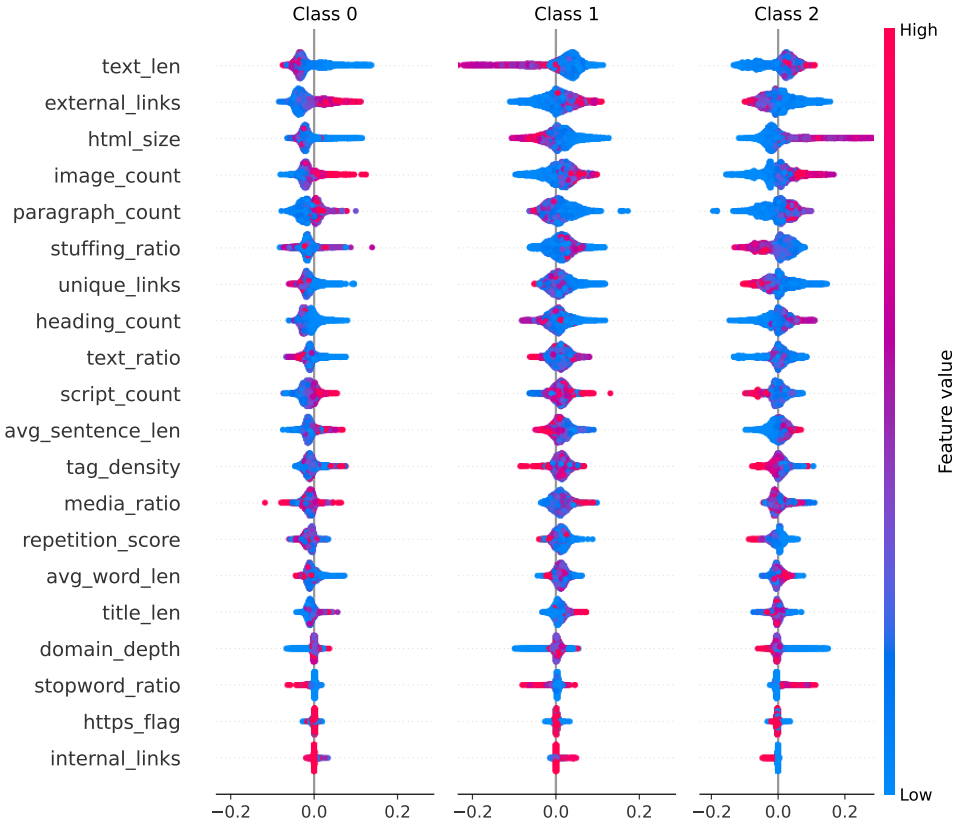


Figure 9: SHAP summary plots for the **Professionalism** across all classes (0, 1, 2).

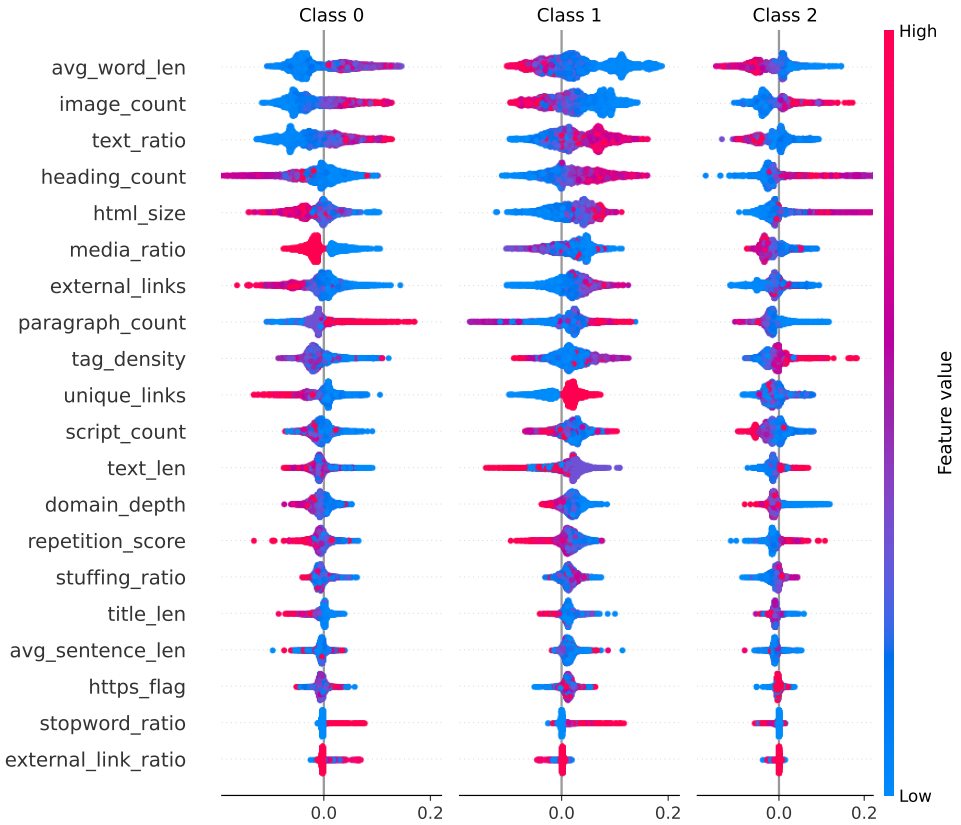


Figure 10: SHAP summary plots for the **Design** across all classes (0, 1, 2).

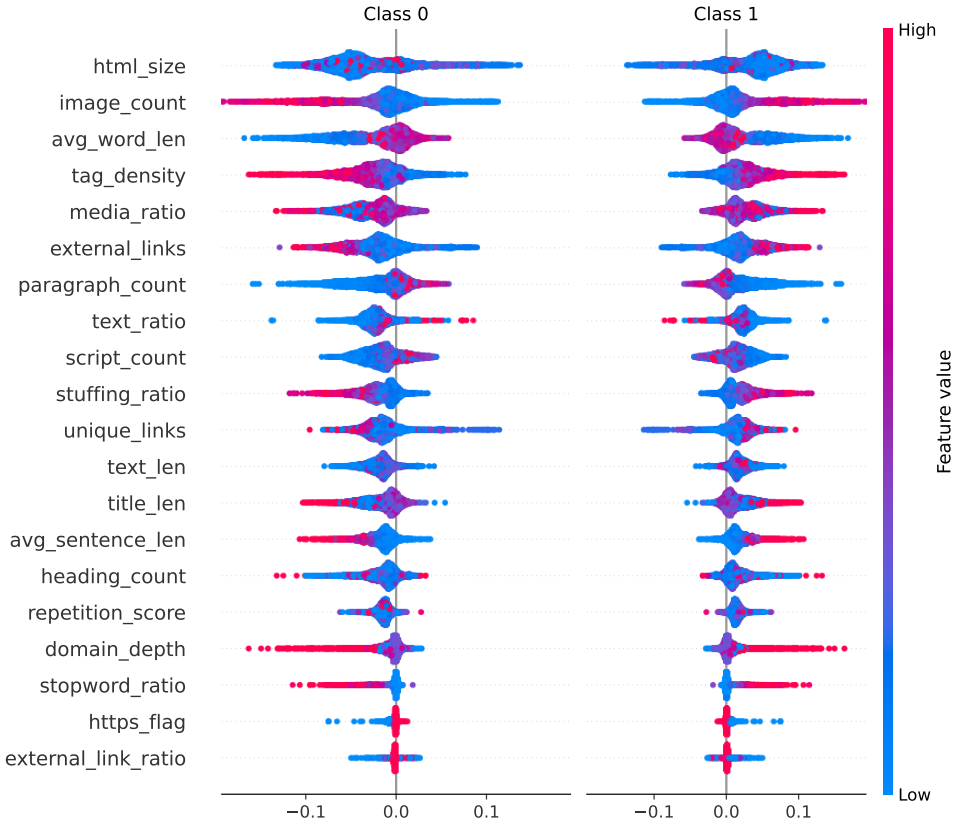


Figure 11: SHAP summary plots for the **Authenticity** across classes (0, 1).


```
    "action_id": 1,
    "feature": "text_len",
    "severity": "high",
    "reasoning": "The page contains only
        12 words, providing
       insufficient content for users
       and search engines.",
    "fix_types": [
      "add_content",
      "expand_body_text"
    ],
    "dependencies": [],
    "description": "Create substantial,
       relevant body copy that covers
       the topic in depth. Aim for at
       least 300-500 words to improve
       readability and SEO. Ensure the
       new text is well-structured with
        paragraphs and subheadings."
  },
  {
    "action_id": 2,
    "feature": "stuffing_ratio",
    "severity": "high",
    "reasoning": "A stuffing_ratio of 1
       indicates extreme keyword
       concentration, which can be
       penalized by search algorithms."
       ,
    "fix_types": [
      "reduce_keyword_density",
      "rewrite_text"
    ],
    "dependencies": [     1     ],
    "description": "Rewrite the newly
       added content to use keywords
       naturally and sparingly.
       Incorporate synonyms and related
        terms to lower the top-10 word
       concentration. Maintain
       readability while preserving
       topical relevance."
  },
  ...]
```


F Robustness Evaluation

This appendix extends the retrieval-shift analysis from Section 5.4 to the remaining four *WebQuality* dimensions: **Relevance**, **Content**, **Design**, and **Authenticity**. The evaluation protocol is identical: for every page, we recompute the full set of RAG-related retrieval features before and after optimization and measure directional change. These results provide a robustness check on whether the degradation patterns observed for *overall quality* generalize across dimensions.

F.1 Page-Level RAG Evaluation

This appendix extends the analysis from Section 5.4 to all five *WebQuality* dimensions. For each target dimension, we compute retrieval-related

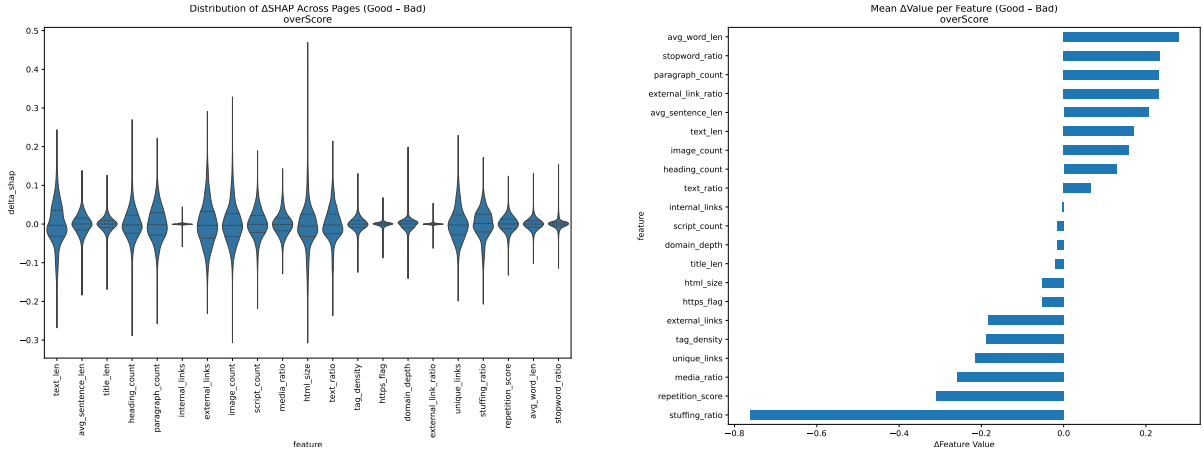


Figure 12: Feature attribution differences for **Overall Quality**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

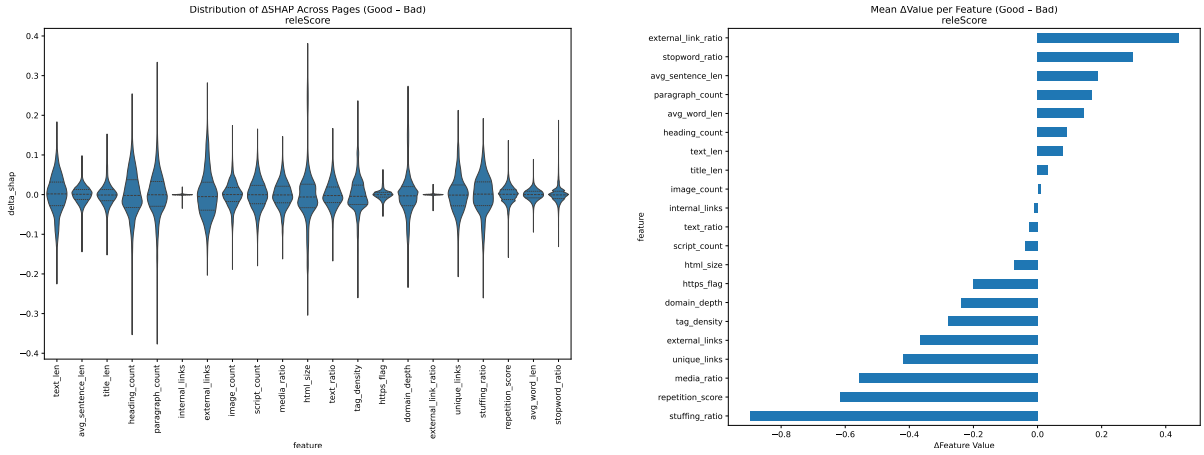


Figure 13: Feature attribution differences for **Relevance**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

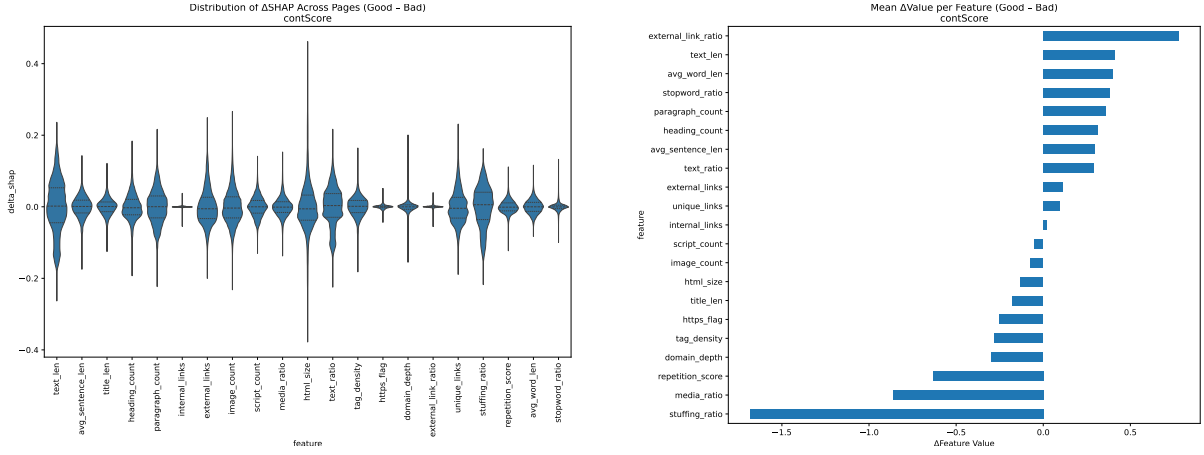


Figure 14: Feature attribution differences for **Content Quality**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

features before and after SHAP-guided optimization, quantifying how human-centered rewriting affects a page's visibility under embedding-based retrieval. Across all dimensions, the dominant pattern persists: optimizations increase structural or surface-level attributes (e.g., `html_chunks`, `html_time`), while semantic similarity features (`query_html_score`, `title_html_score`, cover-

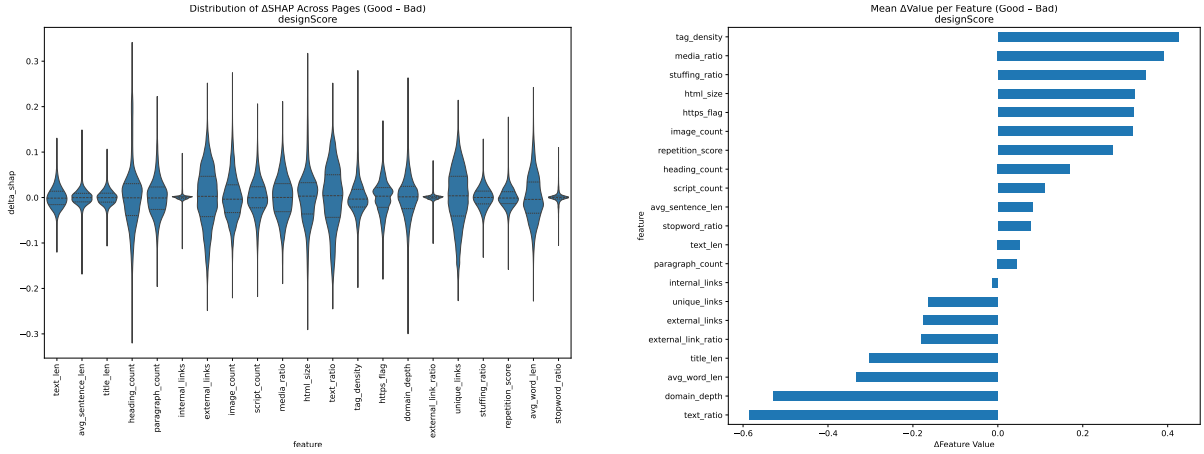


Figure 15: Feature attribution differences for **Design Quality**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

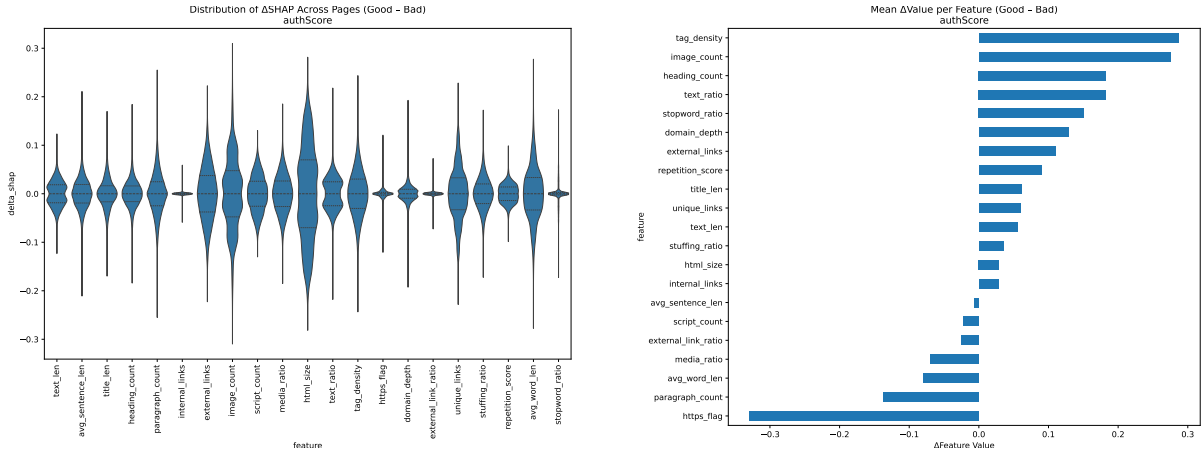


Figure 16: Feature attribution differences for **Authenticity**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

age metrics) consistently degrade. Below, we provide dimension-specific results.

Relevance

Across the rewritten pages, almost all retrieval-oriented indicators degrade. Between 60–86% of pages show declines in semantic alignment features, including `html_score` (63% degraded), `query_html_score` (54%), `title_html_score` (63%), and most coverage measures (41–49%). Even structural proxies such as `html_chunks` and `html_time` degrade for 56% and 86% of pages respectively. Only a minority of features show partial improvement, and those improvements are typically small compared to the scale of degradation. Overall, relevance-focused optimization reliably introduces semantic drift in embedding space, weakening alignment to both the document's original meaning and its associated query (Figure 18a).

Content Quality

For the content-optimized pages, degradation is even more pronounced. The majority of structural and semantic dimensions decline: `html_chunks` (88% degraded), `html_score` (70%), `query_html_score` (79%), `query_fusion_score` (79%), and all title–HTML similarity measures (55–76%). A few structural features exhibit moderate improvement, but the overall pattern remains overwhelmingly negative. These results indicate that content-enriching rewrites—despite adding text—tend to disrupt embedding consistency and diminish retrieval-aligned signals (Figure 18b).

Design

Design-oriented optimization shows the strongest and most consistent degradation across all dimensions. Among 522 rewritten pages, over 60% degrade on nearly every retrieval fea-

Table 9: Per-dimension comparison of optimization outcomes across three strategies. Values represent the proportion (%) of pages that **improved** (↑), remained **unchanged** (↔), or **degraded** (↓). Each SHAP-guided model optimizes only for the same target dimension used in its SHAP attribution computation.

Dimension	SHAP-guided			No SHAP			No SHAP/No Feats		
	↑	↔	↓	↑	↔	↓	↑	↔	↓
overScore	**73.74**	26.36	0.00	45.65	54.35	0.00	60.05	39.95	0.00
releScore	**74.83**	25.17	0.00	55.1	44.90	0.00	69.39	30.61	0.00
contScore	**96.97**	3.03	0.00	72.73	27.27	0.00	96.97	3.03	0.00
designScore	26.05	73.95	0.00	8.62	91.38	0.00	**31.42**	68.58	0.00
authScore	**86.67**	13.33	0.00	30.00	70.00	0.00	73.33	26.67	0.00

ture: `html_score` (62%), `query_html_score` (62%), `query_fusion_score` (62%), `title_html_score` (65%), and several coverage metrics. Structural indicators also decline: `html_time` degrades for 86% of pages, `html_chunks` for 66%, and `html_spike_ratio` for 61%. The consistency of these drops suggests that design-oriented rewrites introduce markup patterns or phrasing that dense retrievers interpret as noisy or off-distribution, decreasing retrievability (Figure 18c).

Planner Agent Prompt

You are the Optimization Planner Agent. Given the feature explanations, the page-specific feature table, and the reference differences between high- and low-quality pages, produce a JSON list of optimisation actions.
Each action must include:

- `action_id`
- `feature`
- `severity` (`low|medium|high`)
- `reasoning`
- `fix_types` (list)
- `dependencies` (list of IDs)
- `description` (1–3 sentences)

Context Provided:

- Feature explanations
- Page-specific feature values and SHAP impacts
- Reference differences between good and bad pages

Respond only with a JSON array of actions.

Figure 17: Prompt used by the Optimization Planner Agent.

Authenticity

Authenticity-driven rewrites shows mixed but still largely negative outcomes. Structural features such as `html_chunks` (70% degraded) and `html_time` (73%) show substantial declines. Semantic indicators also worsen for most pages: 60–80% degrade in coverage metrics. A few features exhibit mild improvement (e.g., `query_fusion_score`, `query_html_score`, and `html_score`), but these gains are neither consistent nor large enough to offset pervasive degradation. Overall, authenticity-focused edits introduce additional textual detail but reduce embedding alignment and retrieval consistency (Figure 18d).

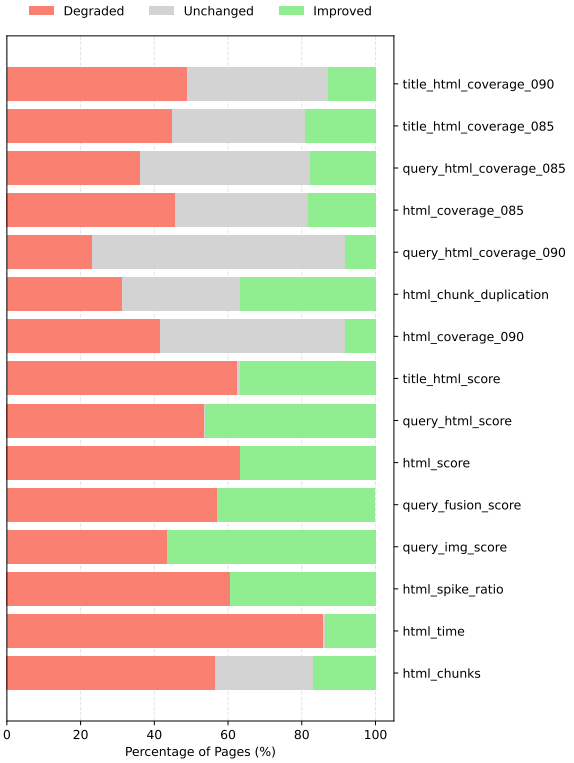
F.2 Multi-Embedder Evaluation

To verify that the retrievability degradation observed in Section 5.4 is not specific to any single dense retriever, we repeat the document-store experiment using three widely deployed embedding models:

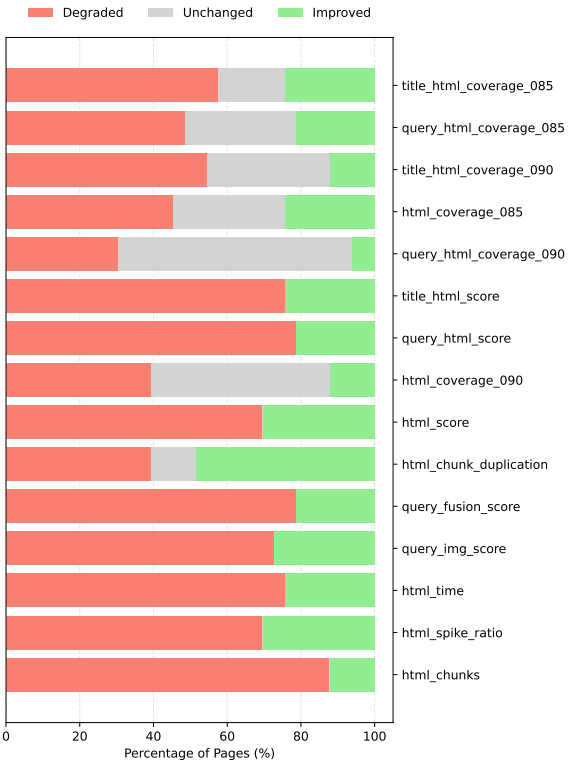
- `BAAI/bge-large-zh` (semantic baseline)
- `intfloat/multilingual-e5-large` (retrieval-tuned)
- `moka-ai/m3e-base` (our main embedder)

For each model, we build two FAISS indexes—one using the original pages and one where low-quality pages are replaced with their optimized versions—and evaluate document-level retrievability for all "bad" pages. The metrics include mean rank, Recall@10, and MRR.

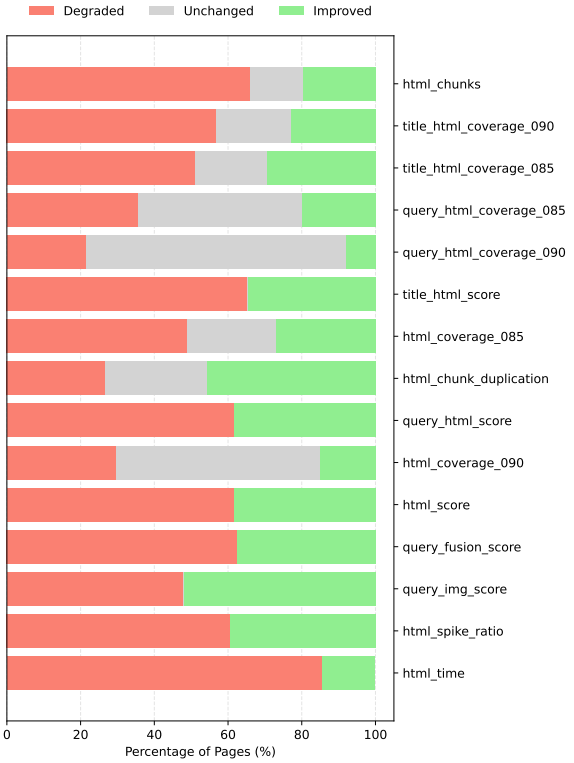
Delta Computation. For each metric we report the difference between the original and optimized retrieval scores. Since lower ranks are better but



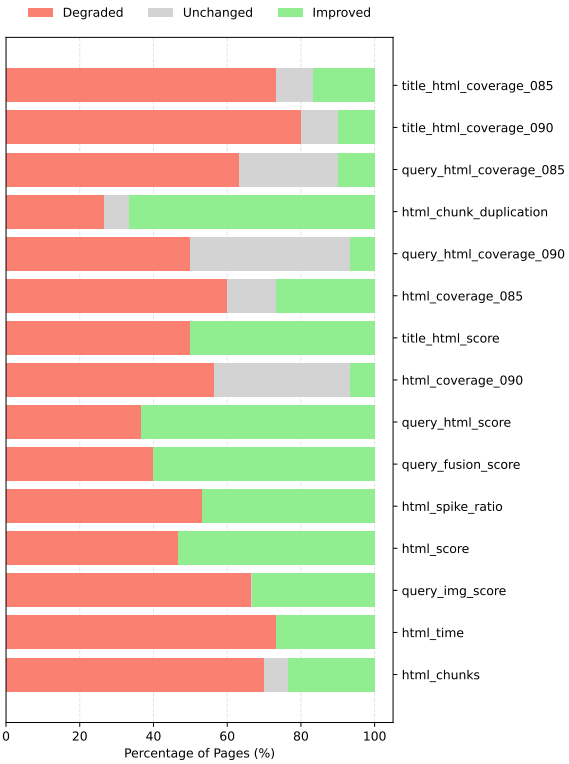
(a) **releScore:** Surface-level features occasionally improve, but semantic similarity degrades for the majority of pages (53–62%).



(b) **contScore:** Structural expansion increases sharply, while semantic features degrade for most pages (57–79%).



(c) **designScore:** Design-oriented rewrites increase visual and structural complexity, while semantic alignment consistently declines.



(d) **authScore:** Authenticity-focused edits add descriptive detail but reduce embedding-level coherence for many pages.

Figure 18: RAG feature shifts after SHAP-guided optimization across the *relevance*, *content*, *design*, and *authenticity* dimensions. Across all four settings, pages become structurally richer but semantically less aligned with their original content and the corresponding queries.

higher Recall@10/MRR values are better, deltas are computed as:

$$\Delta\text{Rank} = \text{Rank}_{\text{orig}} - \text{Rank}_{\text{opt}},$$
$$\Delta\text{Recall@10} = \text{Recall@10}_{\text{orig}} - \text{Recall@10}_{\text{opt}},$$
$$\Delta\text{MRR} = \text{MRR}_{\text{orig}} - \text{MRR}_{\text{opt}}.$$

With this convention, positive values of ΔRank indicate improvement, while negative values indicate degraded retrievability.

For ΔRecall@10 and ΔMRR, the opposite holds: positive values indicate degradation.

Summary of Findings. Across all three embedders and four of the five *WebQuality* dimensions, the optimized pages become *less* retrievable:

- ranks worsen ($\Delta\text{Rank} < 0$),
- Recall@10 decreases,
- and MRR declines accordingly.

The only consistent exception is *authScore*, where the optimization induces nearly no change. For *designScore*, degradation remains present but is somewhat attenuated depending on the embedder, indicating that layout- and structure-heavy edits interact more variably with embedding geometry.

These cross-model results confirm that the retrievability drop is a systematic phenomenon and not an artifact of any single encoder.

Table 10 provides the aggregated deltas per dimension.

Table 10: Delta-based retrievability comparison across embedders. Positive values of ΔRank indicate improvement after HTML optimization, while negative values indicate degraded retrievability. For ΔRecall@10 and ΔMRR, the opposite holds.

Embedder	**Dimension**	**ΔRank**	**ΔRecall@10**	**ΔMRR**
`BAAI/bge-large-zh`	overScore	-0.26	0.05	0.05
	releScore	-0.25	0.05	0.04
	contScore	-0.31	0.04	-0.01
	designScore	-0.39	0.06	0.05
	authScore	-0.08	0.01	-0.01
`intfloat/multilingual-e5-large`	overScore	-0.14	0.05	0.05
	releScore	-0.06	0.05	0.04
	contScore	-0.10	0.04	-0.01
	designScore	0.02	0.06	0.05
	authScore	0.04	0.01	-0.01
`moka-ai/m3e-base` (used in our study)	overScore	-0.33	0.02	0.04
	releScore	-0.45	0.003	0.05
	contScore	-0.39	0.008	0.01
	designScore	-0.25	0.02	0.03
	authScore	0.00	0.00	0.002