# Human-Centred Webpage Optimization Degrades Dense Retrieval

**Anonymous ACL submission**

## Abstract

Webpages increasingly serve two audiences: *humans*, who assess credibility and usefulness, and *machines*, which surface pages in retrieval-augmented generation (RAG) pipelines. Yet it remains unclear whether improving webpages for human readers also improves their visibility to dense retrievers. We introduce WEBQX, an interpretable framework for analysing the relationship between human-perceived webpage quality and retrieval behaviour. Built on the WEBQUALITY dataset of more than 60k annotated webpages, the framework combines quality prediction from structural HTML features, SHAP-guided HTML rewriting, and retrieval evaluation. Our experiments show a consistent tension between human-centred webpage optimization and dense retrieval. While the proposed optimization pipeline reliably improves predicted webpage quality, dense retrieval performance systematically deteriorates. The effect persists in hybrid retrieval settings, whereas purely lexical retrieval remains largely unaffected. These findings provide the first large-scale evidence of a structural misalignment between human-centred webpage optimization and dense retrieval, highlighting the need for jointly optimized retrieval and web-design strategies in RAG-mediated web ecosystems. Code and trained models are publicly available: `https://anonymous.4open.science/r/webqxaisq-B38F/README.md`

## 1 Introduction

RAG systems increasingly rely on external webpages as grounding sources for LLMs (Lewis et al., 2020; Gao et al., 2023). In these systems, dense retrievers effectively determine which webpages become visible to both models and users. At the same time, webpage quality strongly influences user trust, readability, and perceived credibility. However, the structural signals preferred by dense retrieval systems may not align with human notions of webpage quality. This raises an important question: **How does optimizing webpages for human-perceived quality affect their retrievability in dense retrieval pipelines?**

Prior work on web quality assessment has studied readability, credibility, structure, and content organization (Deepthi and Shalini, 2014; Castillo et al., 2011), while recent multimodal approaches such as Hydra learn quality representations directly from webpage content (Zhang et al., 2025). In parallel, retrieval research has investigated ranking robustness and grounding in RAG pipelines (Yu et al., 2024; Gao et al., 2023). Despite progress in both areas, the relationship between human-centred webpage optimization and dense retrievability remains largely unexplored.

To address this gap, we introduce **WEBQX** (Figure 1), an interpretable framework for analysing the interaction between webpage quality and dense retrieval behaviour. Built on the WEBQUALITY dataset (Zhang et al., 2025), the framework consists of three components. WEBQX-ESTIMATOR predicts perceived webpage quality from handcrafted HTML features and provides SHAP-based explanations (Lundberg and Lee, 2017). WEBQX-OPTAGENT uses these explanations to guide targeted HTML rewrites through a two-agent LLM optimization pipeline (Gong et al., 2025). Finally, WEBQX-RAGEVAL evaluates how structural edits affect retrievability across multiple retrieval settings. In summary, our contributions are:

1. WEBQX: a unified framework connecting human-perceived webpage quality and dense retrievability through an interpretable feature space combining HTML structure and retrieval-oriented embedding signals.
2. An interpretable optimization pipeline that performs targeted HTML rewrites and consistently improves predicted human-perceived quality across five quality dimensions.
3. The first large-scale evidence that improving

human-perceived webpage quality can systematically degrade dense retrieval performance in RAG systems.

Our experiments show a consistent tension between human-centred webpage optimization and dense retrieval behaviour. Although the proposed optimization pipeline reliably improves predicted webpage quality, it simultaneously reduces dense retrievability. These findings suggest that dense retrievers are sensitive to structural webpage regularities beyond semantic relevance alone, highlighting the need for jointly optimized retrieval and web-design strategies in RAG-driven 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) 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 retrieval components (e.g., dense and lexical) 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). Further, 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-centred 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 behaviour. 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

Our goal is to analyse how structural HTML changes that improve human-perceived webpage quality affect dense retrieval behaviour. To this end, we first construct an interpretable feature space that captures both human-centred webpage characteristics and retrieval-oriented signals (Sec. 3.1). Based on this representation, we introduce WEBQX, a three-component framework consisting of the WEBQX-ESTIMATOR, which predicts human-perceived webpage quality and generates SHAP-based feature attributions (Sec. 3.2); WEBQX-OPTAGENT, a two-agent LLM pipeline that uses these explanations to perform targeted HTML rewrites (Sec. 3.3); WEBQX-RAGEVAL, which evaluates how these edits affect retrievability across dense, lexical, hybrid, and reranking-based retrieval settings (Sec. 3.4).

### 3.1 Interpretable Feature Space

We construct a feature space comprising two groups of indicators: (1) handcrafted HTML features designed to capture human-perceived webpage quality, and (2) retrieval-oriented metrics that approximate how retrievers score and rank pages. These complementary feature families reflect both, the structural organization of a webpage and its retrieval-facing semantic footprint. We use this feature space to test whether dense retrievers exhibit *structural mirroring*, i.e., whether retrieval behaviour is influenced not only by semantic content but also by regularities in HTML structure.

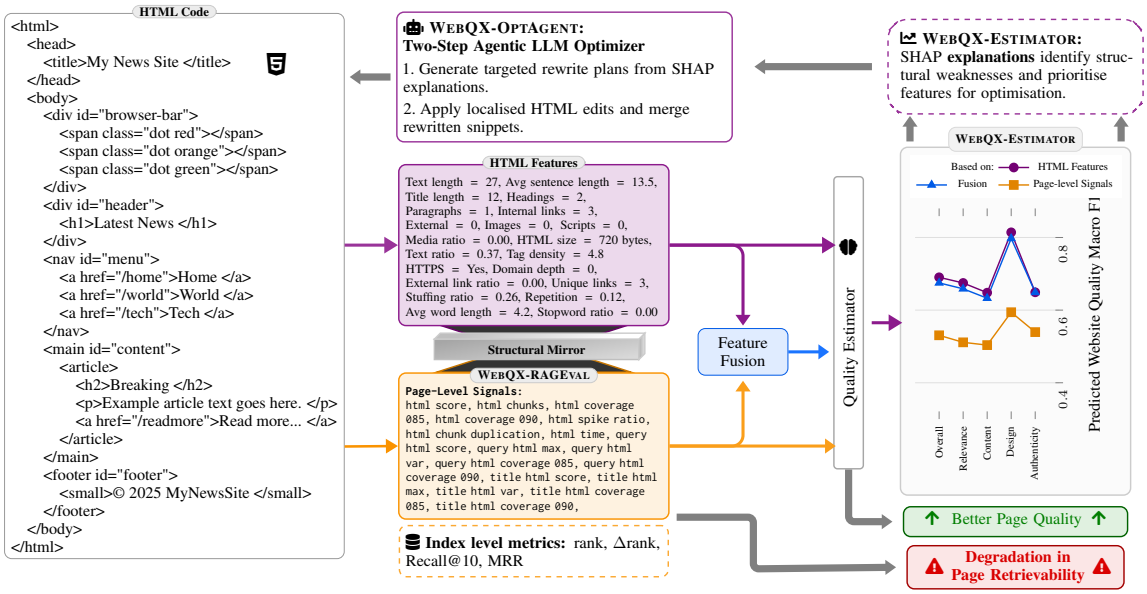



Figure 1: Overview of the proposed WEBQX framework. Webpages are transformed into two complementary feature spaces: interpretable HTML features capturing human-perceived quality, and retrieval-oriented signals derived from dense embeddings. WEBQX-ESTIMATOR predicts webpage quality and produces SHAP explanations identifying structural weaknesses. These explanations guide WEBQX-OPTAGENT, a Two-Step Agentic LLM optimization pipeline that performs targeted HTML rewrites. Finally, WEBQX-RAGEVAL evaluates how the resulting edits affect retrievability at both page and index levels across multiple retrieval settings. The framework reveals a consistent tension between improved human-perceived quality and degraded dense retrievability.

**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 and anchor richness). These features quantify how information is structured, presented, and interlinked – dimensions relevant to both human quality judgments and machine retrievability. Figure 1 shows example values ("HTML Features"), with full feature descriptions in Appendix A.1.

**Retrieval-Driven Page-Level Signals.** To characterize retrieval-relevant signals, we compute embedding-based similarity features aligned with dense retrieval behaviour. 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 et al., 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 dense retrievers identify semantically relevant content. While our broader evaluation includes lexical, hybrid, and reranking-based retrieval, these signals are designed to isolate embedding-level behaviour. The full list of features is shown in Figure 1 ("Pagel-level Signals").

### 3.2 Explainable Web-Quality Prediction

To build the WEBQX-ESTIMATOR, we train classifiers to predict scores for five human-perceived quality dimensions: overall, relevance, content, design, and authenticity. We apply SHAP (Lundberg and Lee, 2017) to each trained model to obtain per-sample, feature attribution scores across all output classes. We aggregate these fine-grained, page-level attributions, to additionally produce global feature importance profiles, highlighting which HTML features most strongly influence predicted human-perceived quality. In summary, we obtain five predicted quality scores for each webpage, along with explanations for each predicted score.

### 3.3 Human-perceived Quality Optimization

To convert explanations into actionable improvements, we first transform the SHAP explanations into a prioritized set of optimization targets. We then use the ESTIMATOR's predictions and SHAP explanations to improve human-perceived webpage quality. Our WEBQX-OPTAGENT comprises two LLM agents: a PLANNER, which translates SHAP feature attributions into structured optimization steps, and an EDITOR, which applies these edits at snippet- or page-level granularity.

**Optimization Targets from Explanations.** Given the raw SHAP 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.

**Agentic Rewriting.** Using SHAP attributions, feature context, and attribution differences as in (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 is shown in Figure 2, with further implementation details in Appendix E.1. For each planned action, the EDITOR 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 (see Appendix E.2).

### 3.4 Retrievability Assessment

Optimized webpages are evaluated with WEBQX-RAGEVAL, which analyses how human-centred HTML edits affect retrieval behaviour. The evaluation is performed at two complementary levels. At the *page level*, WEBQX-RAGEVAL measures how structural edits alter retrieval-oriented representations derived from dense embeddings and document structure. At the *index level*, it evaluates how these representation changes propagate to retrieval performance within dense, lexical, hybrid, and reranking-based pipelines. Together, these analyses enable us to study the relationship between human-perceived webpage quality and retrievability both locally (representation shifts) and globally (retrieval outcomes).

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.

## 4 Experimental Setup

We first introduce the datasets used in this work (Sec. 4.1), followed by the experimental setup for HTML-retrieval correlation analysis (Sec. 4.2), quality prediction (Sec. 4.3), quality optimization (Sec. 4.4), and retrieval evaluation (Sec. 4.5).

### 4.1 Datasets

**WEBQUALITY** (Zhang et al., 2025) contains more than 60,000 Chinese webpages together with their full HTML, extracted text, page titles, and associated user queries. These components enable joint analysis of webpage structure, semantic content, and query-level retrieval behaviour. In addition, WEBQUALITY provides human annotations for five perceived quality dimensions: overall quality, relevance, professionalism, design, and authenticity. Each dimension is rated on an ordinal scale (0: poor, 1: ordinary, 2: excellent). The overall score represents a holistic assessment aggregated across these dimensions. All supervised experiments are performed on WEBQUALITY, since it is the only dataset providing annotations for perceived quality.

To assess whether the observed structural–retrieval relationships generalize beyond a single language, we additionally analyse the multilingual **WEBMMU** corpus (Awal et al., 2025). Similar to WEBQUALITY, WEBMMU provides webpage content and structural information across diverse languages, but does not include perceived quality annotations. We therefore use it to study correlations between structural features and embedding-based retrievability across languages, providing supporting evidence that the observed structural signals are not language-specific.

### 4.2 Feature Analysis

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 Quality Prediction

For WEBQX-ESTIMATOR, we train separate Random Forest classifiers for each quality dimension using z-score normalization, grid-searched hyperparameters, and 10-fold cross-validation (Appendix C). Performance is reported as Macro-F1 ($\pm$ std). As baselines, we include MLPs as a strong tabular neural baseline (Gorishniy et al., 2021), ordinal models (Winship and Mare, 1984) to account for the ordinal labels (poor, ordinary, excellent), and the state-of-the-art multimodal model Hydra (Zhang et al., 2025). Hydra uses three modality-specific encoders whose embeddings are concatenated for prediction. We extend Hydra with our handcrafted HTML features using three integration strategies: (i) projecting the handcrafted features into a latent space with the same dimensionality and concatenating them with Hydra's multimodal embeddings before prediction, (ii) additionally replacing Hydra's linear prediction head with an MLP-based head over the fused representation, and (iii) an attention-based module that estimates adaptive fusion weights.

### 4.4 Quality Optimization

Both the PLANNER and EDITOR agents are implemented using `GPT-OSS-120B`[1]. Optimized pages are subsequently re-scored using ESTIMATOR. We compute $\Delta Q = Q_{\text{after}} - Q_{\text{before}}$ per quality dimension and record quality label changes. For each dimension, we define low-quality webpages as pages with a corresponding dimension-specific quality label of *poor*. To remain within the context limits of `GPT-OSS-120B`, we restrict optimization to pages shorter than 100k characters. Across all dimensions, this results in 1,118 optimization targets.

### 4.5 Retrieval Evaluation

WEBQX-RAGEVAL evaluates retrievability before and after optimization across multiple retrieval settings. At the *page level*, we compare retrieval-oriented features derived from dense embeddings, including embedding similarity, coverage, redundancy, and chunk structure, to quantify representation-level changes. The full feature description is provided in Appendix A.2. At the *index level*, we evaluate dense (FAISS (Douze et al., 2025)), lexical (BM25), hybrid, and reranking-based retrieval pipelines using a known-item retrieval setup, where each webpage is paired with an associated query. Optimized low-quality webpages selectively replace their original representations while the remainder of the corpus remains fixed. We report rank, $\Delta$rank, Recall@10, and Mean Reciprocal Rank (MRR), together with the fraction of webpages whose retrieval performance improves, degrades, or remains unchanged. High-quality webpages serve as an empirical upper bound for retrieval performance.

## 5 Results

We first examine the relationship between HTML structure and dense retrieval behaviour (Sec. 5.1), showing that retrieval signals systematically correlate with structural webpage properties. We then demonstrate that these structural features enable accurate and interpretable prediction of human-perceived webpage quality (Sec. 5.2), and that SHAP-guided edits reliably improve low-quality webpages (Sec. 5.3). Finally, we analyse how these human-centred optimizations affect retrievability (Sec. 5.4), showing a systematic tension between perceived quality and dense retrieval performance.

### 5.1 HTML Structure and Dense Retrieval

In the WEBQUALITY corpus, features measuring query-to-page and title-to-page alignment in embedding space (e.g., *title_html_max*, *html_sim_max*) correlate strongly with HTML

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

| 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. WEBMMU contains four languages (English, Spanish, German, and French). The full feature descriptions can be found in Appendix A and the full correlation matrices in Appendix B.

properties such as content repetition, external linking, and keyword density ($|r| \approx 0.45$–$0.53$) (Table 1). This suggests that dense retrievers encode not only semantic relevance, but also recurring structural patterns in webpage design.

The same trend appears in the multilingual WEBMMU corpus (Awal et al., 2025), where retrieval signals correlate with hyperlink density, media usage, and text-to-markup balance (*external_link_ratio*, *media_ratio*, *text_ratio*; $|r| \approx 0.30$–$0.40$). This indicates that the relationship between HTML structure and dense retrieval behaviour generalizes across languages.

### 5.2 Interpretable Quality Prediction

WEBQX-ESTIMATOR consistently outperforms all baselines across the five quality dimensions (Table 2) achieving macro-$F_1$ scores between 0.65 and 0.81, compared to scores obtained by Hydra (0.42 to 0.67). These results indicate that handcrafted HTML features provide strong signals for modelling human-perceived webpage quality, outperforming methods based on learned multimodal representations derived from webpage text, images, and retrieval-based features (Appendix A.3).

Beyond predictive performance, SHAP-based feature attributions show how structural and textual properties contribute to individual quality dimensions. These explanations directly support our optimization pipeline by guiding LLM-based rewriting of low-quality pages. Feature attribution analyses and case studies are reported in Appendix D.

### 5.3 SHAP-Guided Quality Optimization

SHAP-guided edits consistently improve the quality dimension across the 1,118 optimized webpages (Table 3). Improvements dominate across all dimensions, except *design* with particularly strong gains for *professionalism* (97.0%) and *authenticity* (86.7%). Most improvements correspond to transitions from poor to ordinary quality (0→1), while direct transitions from poor to excellent (0→2) are less frequent and regressions remain absent. For example, optimizing *relevance* improves the corresponding label for 74.8% of webpages, with 71.43% transitioning from 0→1 and 3.4% from 0→2, while 25.2% remain unchanged (Appendix E.3). The *design* dimension shows comparatively limited responsiveness, with the majority of pages remaining unchanged (74.0%). This is likely because the optimization process primarily targets textual, structural, and semantic features derived from HTML content, while visual layout and stylistic aspects of design are only indirectly affected. Beyond the directly optimized dimension, we additionally measure changes in the remaining quality dimensions of the same page. These cross-dimensional effects are generally smaller but remain positive (Appendix E.3).

**Ablations.** To verify that quality improvements are not specific to the Random Forest estimator, we additionally evaluate optimized webpages using independent MLP-based assessors and ordinal regression models (Appendix E.5). Both confirm consistent quality improvements across dimensions.

We further evaluate the contribution of SHAP guidance using two reduced variants: (i) retaining handcrafted features and reference statistics while removing SHAP attribution signals, and (ii) using only textual quality descriptions. The full SHAP-guided pipeline outperforms both baselines on four of five quality dimensions, indicating that attribution-based guidance improves optimization

| 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 web page quality dimensions. We compare standard ML baselines, the multimodal Hydra model (Zhang et al., 2025), and Hydra variants augmented with handcrafted HTML features. 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).

| Optimized Dimension | Impr. | Unch. | Degr. |
|---|---|---|---|
| Overall Quality ($n = 386$) | 73.6 | 26.4 | 0.0 |
| Relevance ($n = 147$) | 74.8 | 25.2 | 0.0 |
| Professionalism ($n = 33$) | 97.0 | 3.0 | 0.0 |
| Design ($n = 522$) | 26.0 | 74.0 | 0.0 |
| Authenticity ($n = 30$) | 86.7 | 13.3 | 0.0 |

Table 3: Quality changes per optimized target dimension. Values denote percentages among webpages that improved (Impr.), did not change (Unch.), or degraded (Degr.). "Impr." aggregates all upward transitions (0→1 and 0→2).

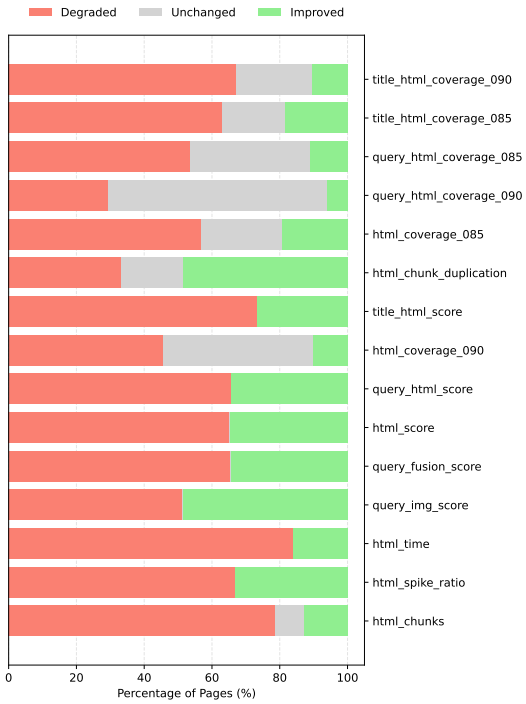



Figure 3: Overall Quality: Distribution of RAG feature shifts after SHAP-guided human-centred optimization. All structural and semantic retrieval features degrade for the majority of pages, indicating that edits that improve predicted human quality often reduce dense retriever compatibility. A few features,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

outcomes. The only exception is again the quality dimension *design*, suggesting that visual design cues are insufficiently captured by HTML-level structural features (full details in Appendix E.4).

### 5.4 Retrieval Effects of Quality Optimization

**Page-level effects.** Most retrieval-oriented features decline after optimization (Figure 3 for overall quality, Appendix F.1 for all quality dimensions). Signals related to content segmentation and structural regularity (e.g., chunk count, processing-time proxies, and spike ratios) decrease for 67–84% of webpages, indicating reduced alignment with retriever-preferred document structures. Measures of query-document semantic alignment (e.g., page-, title-, and fusion-level similarity scores) also decline for 65–73% of webpages. Only a small subset of features remains stable or improves after optimization. For example, high-threshold coverage remains unchanged for 46% of webpages, while chunk duplication improves for 49%. Overall, these trends suggest a tension between human-centred HTML optimization and dense retrieval-oriented representations.

**Index-level effects.** Optimized webpages exhibit consistently lower retrievability in dense retrieval systems (Table 4, Dense). Mean rank worsens (up to $\Delta\text{Rank} = 0.44$), Recall@10 decreases by approximately 2–3 points, and degraded webpages (5–18%) outnumber improved ones (5–8%). In contrast, high-quality webpages remain comparatively stable (mean rank $\approx 3.1$, Recall@10 $\approx 0.79$), suggesting that optimization disrupts retrieval-relevant structural cues while improving perceived quality. Similar patterns are observed across additional embedding models (Appendix F.2).

| | | Mean Rank ↓ | | | ΔRank ↓ | Recall@10 ↑ | | | MRR ↑ | | | Retrieval Change | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Dimension** | Good | Orig | Opt | | Good | Orig | Opt | Good | Orig | Opt | Improved | Degraded |
| Dense | Overall | 3.07 | 2.97 | 3.25 | 0.28 | 0.789 | 0.762 | 0.741 | 0.512 | 0.555 | 0.516 | 6.1 | 15.0 |
| | Relevance | 3.24 | 2.88 | 3.32 | 0.44 | 0.767 | 0.717 | 0.714 | 0.494 | 0.561 | 0.506 | 5.5 | 17.7 |
| | Professionalism | 3.04 | 2.70 | 2.99 | 0.29 | 0.789 | 0.683 | 0.675 | 0.521 | 0.586 | 0.571 | 5.1 | 16.5 |
| | Design | 3.12 | 3.53 | 3.77 | 0.24 | 0.800 | 0.749 | 0.723 | 0.510 | 0.454 | 0.420 | 8.6 | 17.2 |
| | Authenticity | 3.25 | 3.01 | 3.05 | 0.04 | 0.776 | 0.704 | 0.704 | 0.494 | 0.486 | 0.484 | 5.4 | 4.5 |
| Hybrid | Overall | 3.37 | 3.57 | 3.72 | 0.15 | 0.067 | 0.085 | 0.066 | 0.031 | 0.034 | 0.026 | 1.8 | 12.7 |
| | Relevance | 3.56 | 3.30 | 3.53 | 0.23 | 0.070 | 0.073 | 0.054 | 0.030 | 0.032 | 0.021 | 6.2 | 0.0 |
| | Professionalism | 3.34 | 3.71 | 3.78 | 0.07 | 0.07 | 0.175 | 0.15 | 0.032 | 0.062 | 0.053 | 0.0 | 5.6 |
| | Design | 3.55 | 3.56 | 3.71 | 0.15 | 0.073 | 0.074 | 0.065 | 0.032 | 0.031 | 0.028 | 9.4 | 15.1 |
| | Authenticity | 3.67 | 3.00 | 3.00 | 0.00 | 0.076 | 0.044 | 0.031 | 0.032 | 0.019 | 0.015 | 0.0 | 20.0 |
| Reranked | Overall | 2.74 | 2.68 | 2.74 | 0.06 | 0.042 | 0.029 | 0.026 | 0.027 | 0.016 | 0.015 | 15.0 | 15.0 |
| | Relevance | 2.83 | 3.20 | 3.25 | 0.05 | 0.039 | 0.032 | 0.025 | 0.023 | 0.016 | 0.013 | 12.5 | 12.5 |
| | Professionalism | 2.81 | 3.00 | 3.00 | 0.00 | 0.039 | 0.017 | 0.017 | 0.024 | 0.010 | 0.010 | 0.0 | 0.0 |
| | Design | 2.87 | 3.11 | 3.38 | 0.27 | 0.040 | 0.053 | 0.053 | 0.024 | 0.029 | 0.027 | 7.1 | 16.17 |
| | Authenticity | 3.03 | 2.50 | 1.67 | -0.83 | 0.045 | 0.025 | 0.019 | 0.025 | 0.014 | 0.013 | 0.0 | 0.0 |
| BM25 | Overall | 2.20 | 1.86 | 2.10 | 0.24 | 0.048 | 0.032 | 0.034 | 0.034 | 0.023 | 0.022 | 8.0 | 20.0 |
| | Relevance | 2.36 | 2.00 | 1.75 | -0.25 | 0.048 | 0.032 | 0.025 | 0.032 | 0.022 | 0.018 | 0.0 | 0.0 |
| | Professionalism | 2.21 | 1.67 | 2.00 | 0.33 | 0.046 | 0.025 | 0.025 | 0.033 | 0.019 | 0.015 | 0.0 | 33.3 |
| | Design | 2.03 | 2.56 | 2.42 | -0.14 | 0.050 | 0.063 | 0.062 | 0.036 | 0.040 | 0.039 | 5.9 | 7.8 |
| | Authenticity | 2.35 | 1.00 | 1.33 | 0.33 | 0.057 | 0.025 | 0.019 | 0.037 | 0.025 | 0.016 | 0.0 | 33.3 |

Table 4: Retrievability before and after optimization across quality dimensions. Low-quality webpages are optimized; high-quality webpages serve only as a reference group and are not edited. Arrows indicate whether lower (↓) or higher (↑) values correspond to better retrieval performance. ΔRank denotes the mean rank change from original to optimized webpages. Improved and Degraded indicate the percentage of optimized webpages whose retrieval rank improved or worsened.

Despite our focus on dense retrieval to isolate any other effects, we additonally evaluate lexical (BM25), hybrid (BM25 + dense via reciprocal rank fusion), and a two-stage pipeline with cross-encoder reranking. The same directional trend as with dense retrieval persists across all additional settings, although the effect becomes weaker. For instance, ΔRank always worsens in hybrid, but at most by 0.23, whereas in reranking and BM25 ΔRank partially worsens and partially improves. While Recall@10 changes by up to approximately 3 points in dense and hybrid, changes almost vanish under reranking and BM25 (change at most by 0.6 points). Similarly, MRR degrades most for dense (by 0.029 on average), less for hybrid (0.007 on average), and least for reranking and BM25 (below 0.004). On average, the percentage of optimized webpages whose retrieval rank degraded is higher than those whose rank improved in all settings. For BM25, we attribute this to initially high ranks though (e.g., Mean Rank of 1.00 in last row "Authenticity" can not improve, but only degrade), suggesting that the observed degradation is specific to dense semantic representations rather than a universal property of retrieval systems. These findings demonstrate that dense retrievers partially rely on structural regularities that are weakened by edits improving human-perceived webpage quality.

## 6 Conclusion

This work provides the first empirical evidence that improving human-perceived webpage quality can unintentionally reduce retrievability in dense retrieval systems. Across multiple datasets and retrieval settings, we show that dense retrievers systematically encode structural webpage regularities beyond semantic relevance alone. While optimizing webpages for human-perceived quality reliably improves user-facing quality signals, it simultaneously weakens retrieval-oriented signals, leading to measurable degradation in dense retrieval performance.

Our findings indicate a systematic tension between human-centred webpage optimization and machine-oriented retrieval behaviour. More broadly, they suggest that future retrieval models and web optimization frameworks should jointly account for readability, structural quality, and retrievability rather than treating them as independent objectives. We hope this work motivates future research on alignment-aware retrieval systems and jointly optimized Web 4.0 infrastructures.

## Limitations

Our analysis is based on static HTML snapshots and therefore does not capture dynamic or JavaScript-rendered webpage behaviour. In addition, the proposed feature space primarily relies on structural, retrieval-derived, and hand-crafted HTML representations, which may not fully capture higher-level visual, or multimodal webpage characteristics accessible to large transformer-based models.

Optimization is further restricted to webpages below a context-length threshold (100k characters), introducing a bias toward shorter and structurally simpler pages. Moreover, we evaluate only HTML markup and do not assess the rendered visual appearance, usability, or accessibility of rewritten webpages. A direct human evaluation of optimized webpages would provide a stronger validation signal for perceived quality improvements.

Moreover, the notion of webpage quality studied here follows the WEBQUALITY annotation framework and does not explicitly capture task-specific usefulness or interaction efficiency. Our findings should therefore be interpreted within this scope.

Finally, our supervised experiments rely primarily on the predominantly Chinese WEBQUALITY dataset. Although analyses on the multilingual WEBMMU corpus suggest that several structural trends generalize across languages, broader multilingual and domain-diverse evaluation is required to establish external validity. Future work should additionally investigate methods for jointly optimizing webpages for both human-perceived quality and retrieval effectiveness.

## 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. 2025. The FAISS Library . *IEEE Transactions on Big Data*, pages 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.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *CoRR*, abs/2106.11959.

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.

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.

Yuxin Wang, Qingxuan Sun, and Sicheng He. 2023. M3E: Moka Massive Mixed Embedding model. huggingface.co. Hugging Face Repository.

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.

## 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.

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

**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. Dimensions design and authenticity are predicted most reliably (acc. $\geq 0.88$), 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 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).

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

| Features | Overall Quality | Relevance | Professionalism | Design | Authenticity |
|---|---|---|---|---|---|
| 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 |

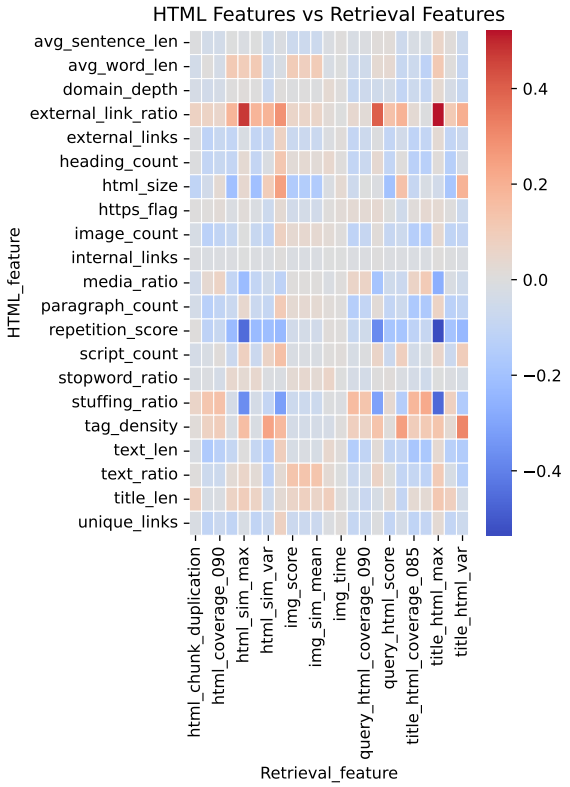



Figure 5: Full Pearson correlation matrix between HTML features and retrieval features in the WEBQUALITY dataset.

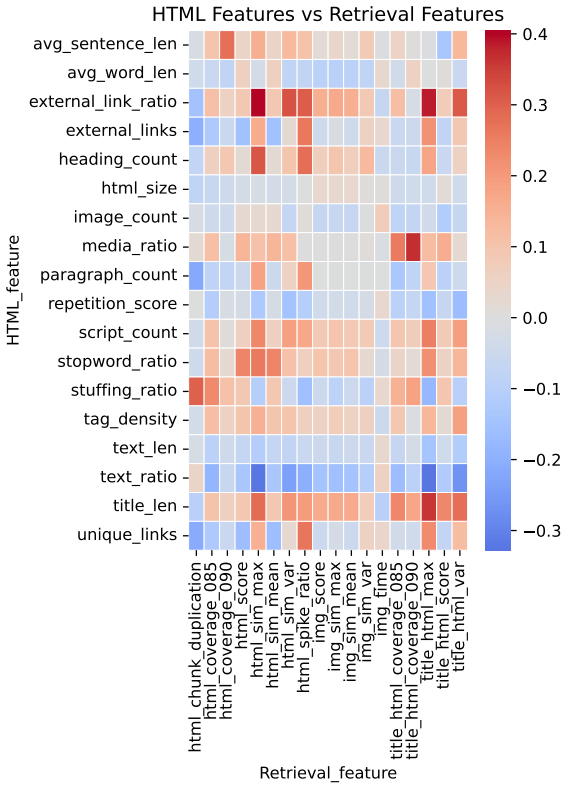



Figure 6: Full Pearson correlation matrix between HTML features and retrieval features in the WEBMMU dataset.

### 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).

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

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. |

this picture. For *relevance* and *content*, natural 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 weak-

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. |

nesses 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 ($\Delta$SHAP and $\Delta$value) 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.

**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.

```
[
  {
    "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."
  },
  ...]
```

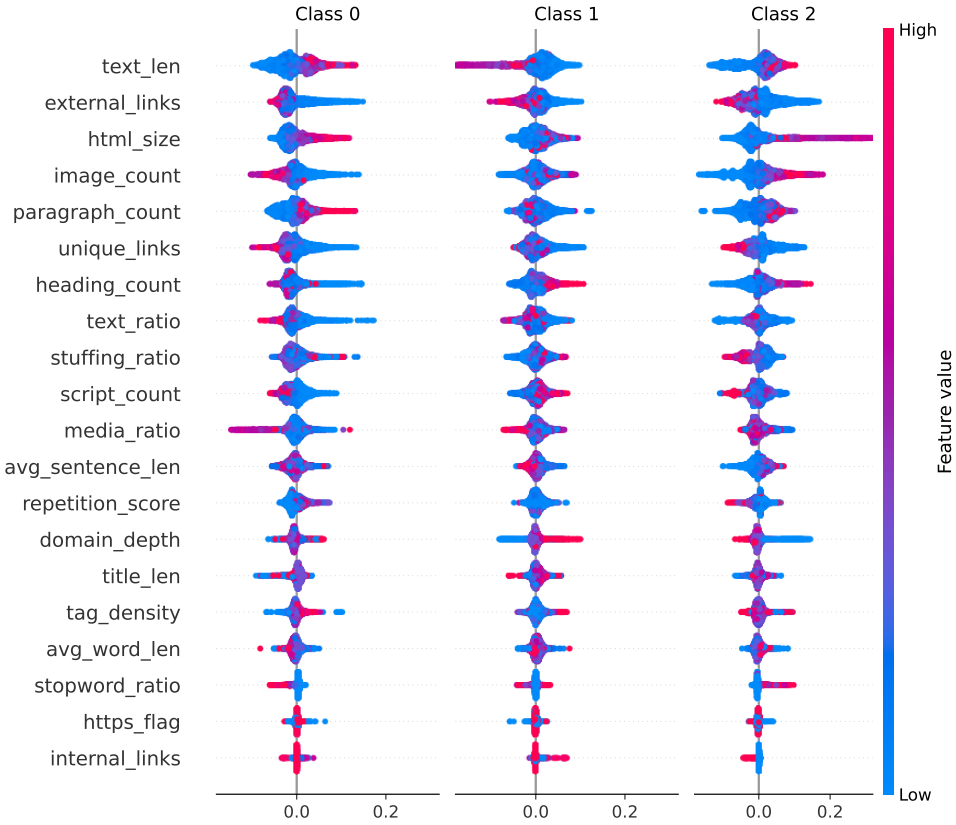



Figure 7: SHAP summary plots for the **Overall Quality** across all classes (0, 1, 2).

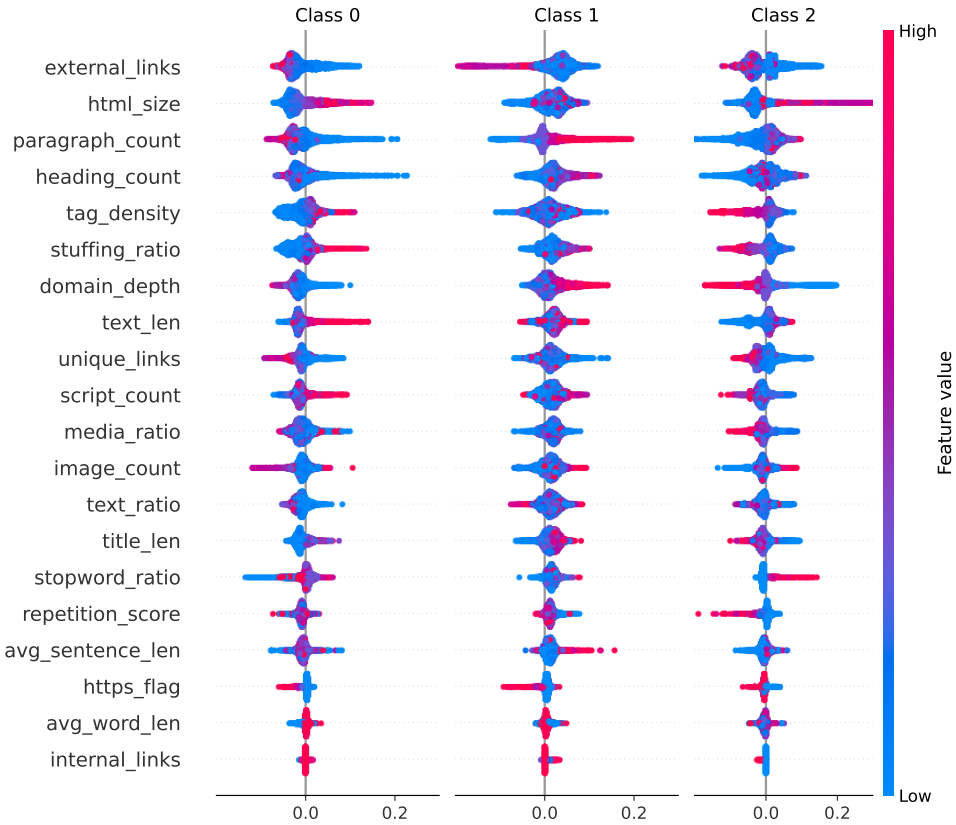



Figure 8: SHAP summary plots for the **Relevance** across all classes (0, 1, 2).

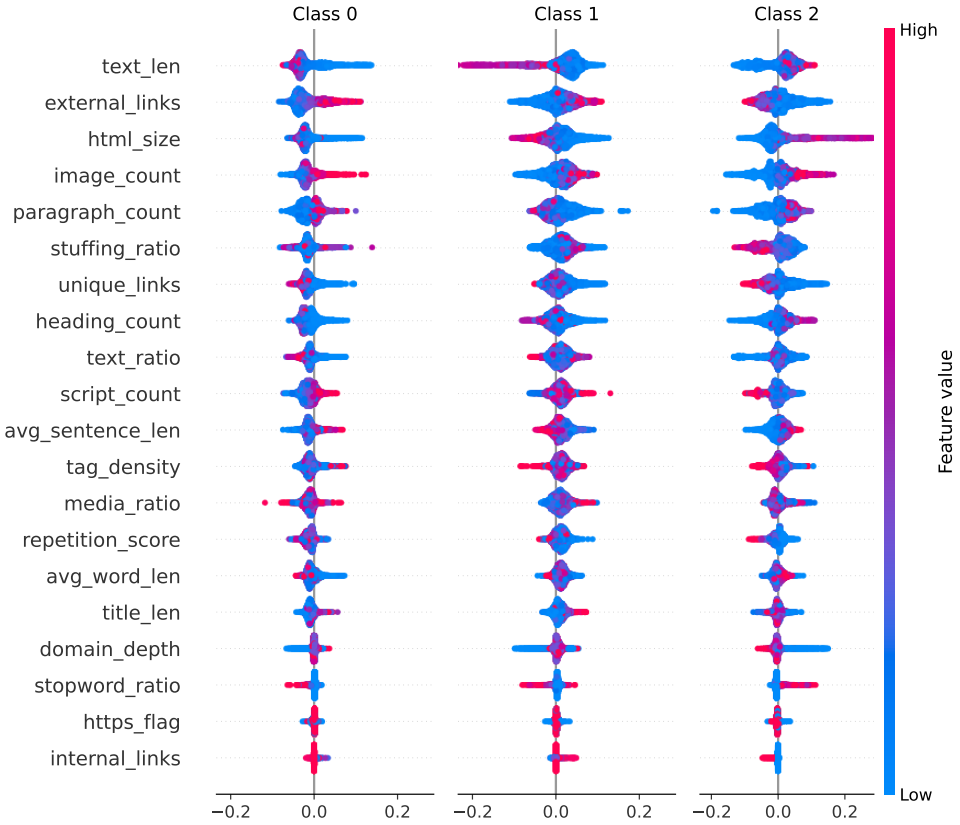



Figure 9: SHAP summary plots for the **Professionalism** across all classes (0, 1, 2).

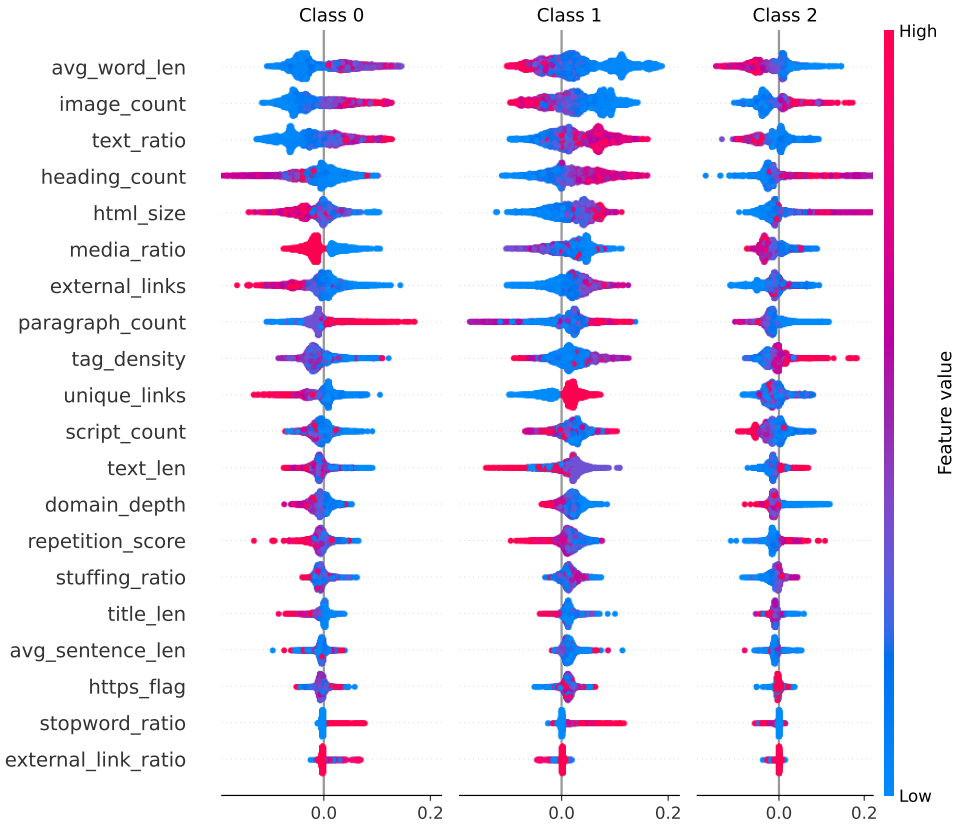



Figure 10: SHAP summary plots for the **Design** across all classes (0, 1, 2).

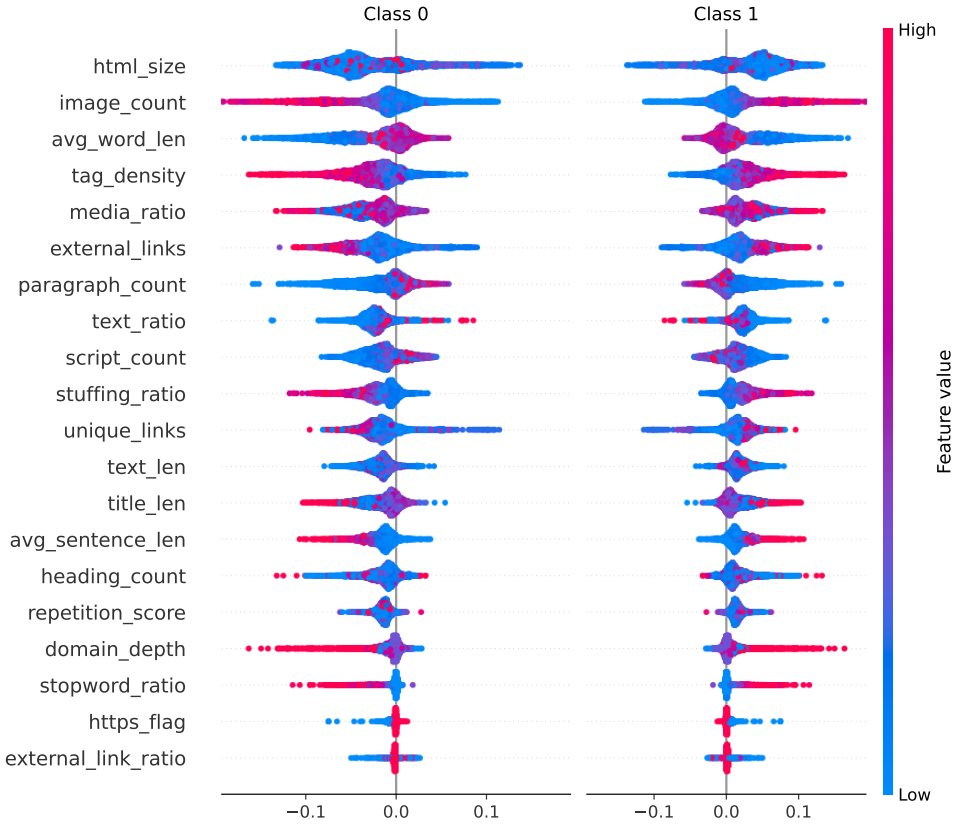



Figure 11: SHAP summary plots for the **Authenticity** across classes (0, 1).

### E.2 Editor Agent

The Editor Agent forms the second stage of the optimisation pipeline. While the Planner Agent determines *what* should be changed and *why*, the Editor Agent is responsible for executing those instructions directly at the HTML level.

Its role is therefore operational rather than analytical: given a structured optimisation action, the agent produces a modified HTML fragment (or full document) that implements the requested change while preserving syntactic validity and overall page coherence.

**Prompt Construction.** For each optimisation action, the Editor Agent receives three inputs:

1. **Full optimisation plan**: the complete ordered list of actions generated by the Planner Agent. This provides global context and allows the agent to reason about dependencies and previously applied edits.

2. **Target action**: a single optimisation instruction specifying the feature to modify, its severity, and the intended fix strategy.

3. **HTML context**: either the full page HTML or a feature-relevant snippet extracted from the document.

The agent is explicitly instructed to apply *only one action at a time*, assuming that all declared dependencies have already been completed. This sequential design prevents conflicting edits and encourages locally consistent modifications. The complete prompt is shown in Figure 18.

**Snippet Selection Strategy.** To minimise unnecessary document rewrites and reduce LLM context requirements, the Editor Agent operates on feature-relevant HTML fragments whenever possible.

Features are divided into two categories:

- **Page-level features** (e.g., `text_len`, `text_ratio`), which require access to the full document.

- **Local structural features** (e.g., `heading_count`, `image_count`), which can be addressed through targeted edits.

For local features, candidate snippets are extracted using lightweight HTML parsing. The most relevant fragment is then selected via semantic similarity between the action description and snippet

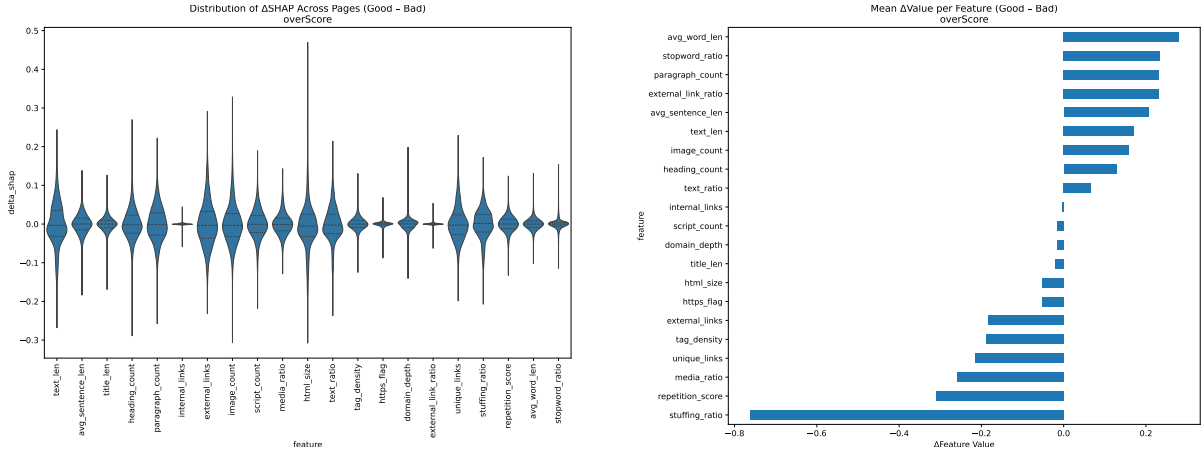



Figure 12: Feature attribution differences for **Overall Quality**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

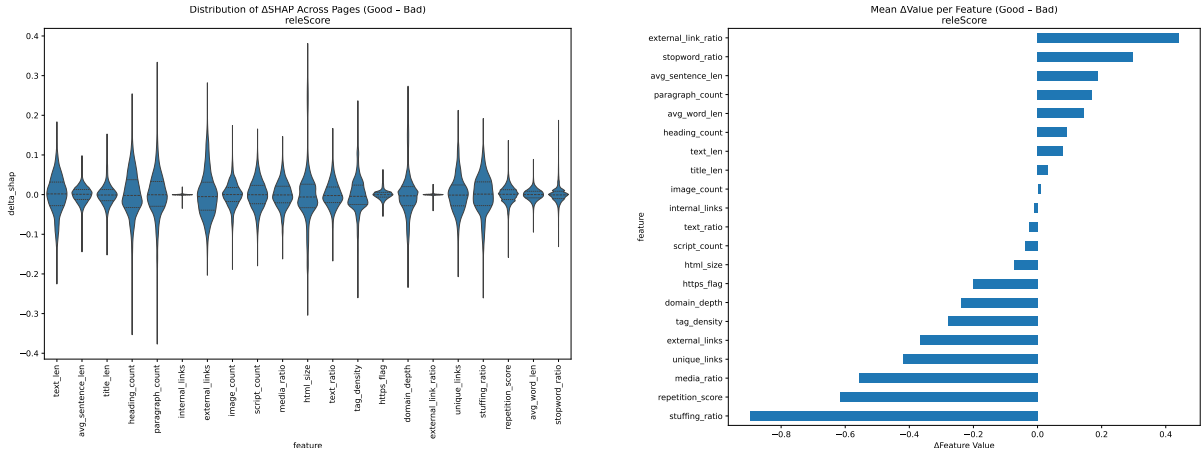



Figure 13: Feature attribution differences for **Relevance**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

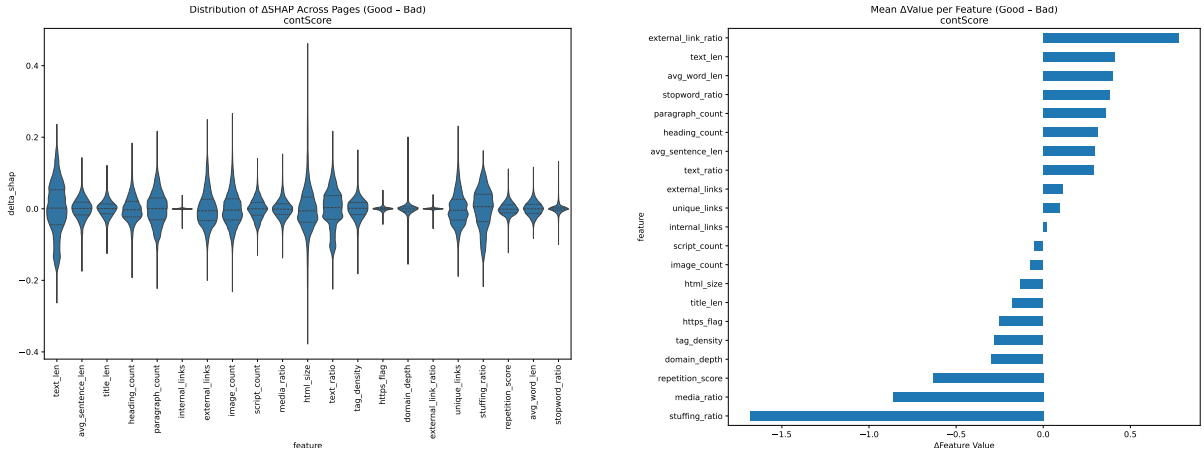



Figure 14: Feature attribution differences for **Content Quality**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

embeddings. This mechanism approximates a retrieval step within the editing process, allowing the agent to focus on the HTML regions most likely to contain the feature-related deficiency.

**Execution Behaviour.** Given the selected context and action specification, the agent produces:

- a modified HTML fragment (or full document), and

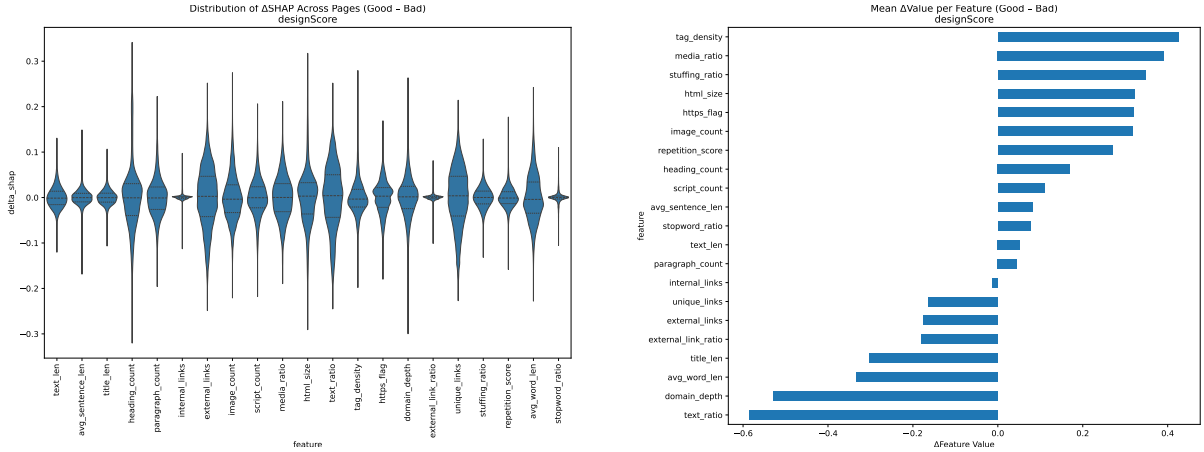



Figure 15: Feature attribution differences for **Design Quality**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

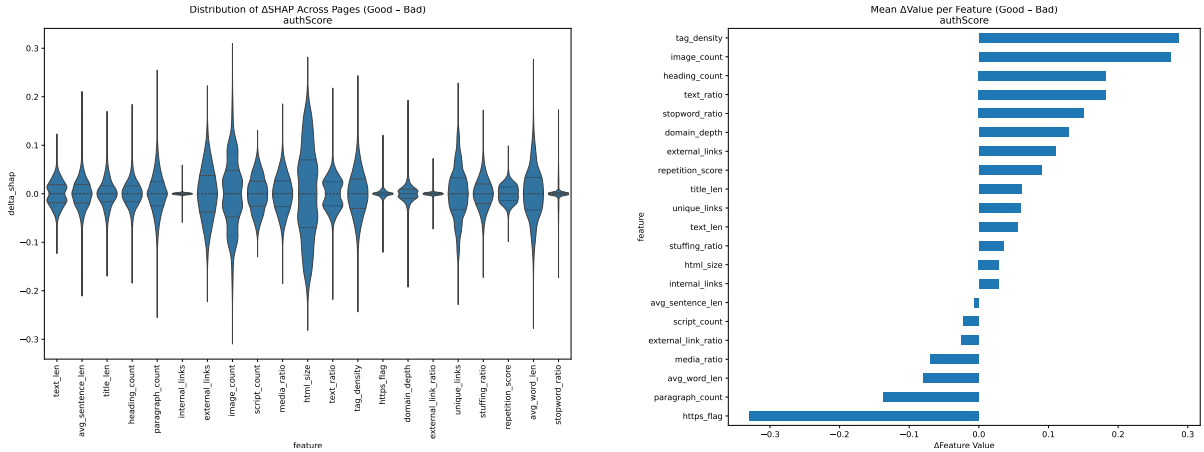



Figure 16: Feature attribution differences for **Authenticity**. Left: distribution of ΔSHAP values (good–bad). Right: mean differences in feature values (good–bad).

- a changelog describing the applied transformation.

The agent is not instructed to optimise stylistic quality or visual aesthetics directly; instead, it performs structural modifications consistent with the Planner Agent's feature-level objectives. This separation preserves the interpretability of the pipeline by ensuring that each edit can be traced back to a specific feature-driven instruction.

**Design Rationale.** The two-agent decomposition introduces a deliberate separation between diagnostic reasoning and HTML manipulation.

This design provides three practical advantages:

1. **Interpretability preservation**: optimisation logic is explicitly represented in the Planner Agent's structured actions rather than embedded implicitly within a single LLM prompt.
2. **Edit locality**: the Editor Agent can operate on small HTML fragments, reducing context noise and improving syntactic stability.
3. **Controlled transformations**: sequential action execution prevents large, uncontrolled document rewrites that could obscure the causal relationship between features and retrievability effects.

Together, these mechanisms ensure that HTML modifications remain traceable, modular, and analytically aligned with the SHAP-derived diagnostics driving the optimisation process.

### E.3 Cross-Dimensional Quality Effects

In addition to improvements in the directly optimized dimension, we analyze how targeted edits affect the remaining quality dimensions of the same webpage. Table 9 reports cross-dimensional quality transitions, where each block corresponds to

| Target | Change Type | Overall Quality | Relevance | Professionalism | Design | Authenticity |
|---|---|---|---|---|---|---|
| **Overall Quality** | 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 |
| ($n$=368) | Unchanged | **26.36** | 45.92 | 64.95 | 82.34 | 91.85 |
| | ↓ | **0.00** | 1.36 | 0.00 | 9.24 | 1.90 |
| **Relevance** | 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 |
| ($n$=147) | Unchanged | 35.37 | **25.17** | 82.99 | 93.20 | 99.32 |
| | ↓ | 0.00 | **0.00** | 0.00 | 5.44 | 0.68 |
| **Professionalism** | 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 |
| ($n$=33) | Unchanged | 12.12 | 93.94 | **3.03** | 78.79 | 100.00 |
| | ↓ | 0.00 | 3.03 | **0.00** | 15.15 | 0.00 |
| **Design** | 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 |
| ($n$=522) | Unchanged | 74.33 | 96.93 | 94.44 | **73.95** | 97.89 |
| | ↓ | 6.51 | 0.38 | 0.19 | **0.00** | 0.38 |
| **Authenticity** | 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 |
| ($n$=30) | Unchanged | 43.33 | 96.67 | 86.67 | 73.33 | **13.33** |
| | ↓ | 6.67 | 0.00 | 0.00 | 6.67 | **0.00** |

Table 9: 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 in the test dataset).

pages optimized for a specific target dimension and columns capture changes in all other dimensions.

Overall, cross-dimensional effects are smaller than direct improvements but remain predominantly positive. Optimizing *professionalism*, for instance, yields substantial gains in overall quality (87.88%) and moderate spillover effects on other dimensions. Similarly, optimizing *relevance* improves overall quality for 64.63% of pages, indicating that content-focused edits often generalize beyond the target dimension.

In contrast, *design* exhibits limited cross-dimensional interaction: improvements remain largely confined to the target dimension, and most other dimensions remain unchanged. This suggests that structural or visual edits are less likely to influence content- or credibility-related aspects of quality.

Across all settings, degradations in non-target dimensions are rare, indicating that targeted optimization does not introduce substantial trade-offs between quality dimensions. These findings highlight that SHAP-guided edits primarily yield localized improvements, while still enabling modest positive spillover effects across related aspects of webpage quality.

### E.4 SHAP Ablation Settings

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. As reported in Table 10 SHAP-guided optimizations consistently improves performance across most dimensions. For *overScore*, *releScore*, and *authScore*, SHAP guidance yields substantially higher improvement rates compared to both degraded variants, highlighting the importance of attribution signals beyond raw feature values. The largest relative gain is observed for *authScore*, where removing SHAP explanations leads to a sharp drop in improvement. For *contScore*, both

SHAP-guided and No-SHAP-No-Feature settings achieve equally high gains, suggesting that content-related optimisation is naturally aligned with the generative strengths of the model. In contrast, *designScore* shows only modest improvements across all settings, with the feature-agnostic variant performing slightly better, indicating that design adjustments may benefit less from feature-level attribution. Notably, no degradations are observed in any configuration, suggesting that the optimisation process remains stable even under reduced guidance.

### E.5 Cross-Dimensional Quality Transitions Under Independent Evaluation Models

The main paper evaluates optimization effects using the same Random Forest (RF) model that provides SHAP-based guidance, which raises the possibility of self-reinforcing evaluation effects. To assess whether the observed improvements generalize beyond the guiding estimator, this section analyzes cross-dimensional quality transitions under three independent evaluator groups: multilayer perceptrons (MLP) and ordinal classification models.

The MLP and ordinal models are included as alternative predictive estimators alongside WEBQX-ESTIMATOR (Section 5.3). For each quality dimension, five independent models are trained. None of these models are used to generate optimization plans. While the MLP and ordinal models are trained on the same feature space as WEBQX-ESTIMATOR (see Section 3.2), they differ architecturally and are not involved in SHAP attribution or rewriting.

**Page Selection and Optimization.** For each target dimension, we select pages labeled as low quality by human annotators and correctly identified as low quality by the RF model. SHAP attributions derived from the RF guide structural HTML modifications. The RF thus provides feature-level explanations only and is not used for evaluation in the analyses reported in this section.

**Independent Classifier Evaluation.** Tables 11 and 12 report cross-dimensional transition matrices under the MLP and ordinal evaluators, respectively. Transitions are computed as categorical shifts between the original and optimized versions (0→1, 1→2, 0→2, unchanged, or degradation).

**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.

**Editor Agent Prompt**

*You are the Optimization Executor Agent. Given the full optimisation plan, a target action, and the HTML to modify, apply only the specified action.*
Constraints:

- Apply **only** the provided action
- Assume dependencies are already satisfied
- Preserve valid HTML structure
- Avoid unrelated modifications

**Inputs Provided:**

- Full optimisation plan (ordered actions)
- Target optimisation action
- HTML context (snippet or full page)

**Return JSON with:**

- `modified_html`
- `changelog`

Figure 18: Prompt used by the Optimization Editor Agent.

Table 10: 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** | | |
|---|---|---|---|---|---|---|---|---|---|
| | ↑ | ↔ | ↓ | ↑ | ↔ | ↓ | ↑ | ↔ | ↓ |
| Overall Quality | **73.74** | 26.36 | 0.00 | 45.65 | 54.35 | 0.00 | 60.05 | 39.95 | 0.00 |
| Relevance | **74.83** | 25.17 | 0.00 | 55.1 | 44.90 | 0.00 | 69.39 | 30.61 | 0.00 |
| Professionalism | **96.97** | 3.03 | 0.00 | 72.73 | 27.27 | 0.00 | 96.97 | 3.03 | 0.00 |
| Design | 26.05 | 73.95 | 0.00 | 8.62 | 91.38 | 0.00 | **31.42** | 68.58 | 0.00 |
| Authenticity | **86.67** | 13.33 | 0.00 | 30.00 | 70.00 | 0.00 | 73.33 | 26.67 | 0.00 |

Across target dimensions, both evaluators confirm substantial within-dimension gains. For example, when optimizing *overall quality*, the ordinal model reports 93.2% upward transitions (0→1 or 0→2), compared to 78.5% under the MLP. Similarly, for the *relevance*, improvements reach 88.4% (ordinal) and 83.7% (MLP). For the *professionalism*, improvements exceed 87% under both models. Even for structurally more challenging dimensions such as the *design*, the ordinal model reports 60.7% improvement (vs. 41.4% under the MLP). For the *authenticity*, improvements reach 96.7% (ordinal) and 100% (MLP), noting the smaller sample size.

Cross-dimensional spillovers are generally positive and moderate. For instance, optimizing the *overall quality* also improves the *relevance* in over 55% of cases under both evaluators, while degradations remain rare and typically below 4%. The consistency across two architecturally distinct classifiers suggests that improvements are not confined to a single model's decision boundary.

Taken together, Tables 11 and 12 demonstrate that SHAP-guided structural optimization yields consistent improvements across multiple independent evaluators, while also revealing meaningful trade-offs in presentation-sensitive dimensions.

# 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 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`, coverage 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 19a).

### 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

Table 11: Cross-dimensional quality transition matrix after SHAP-guided HTML optimization. For each primary dimension (left), pages were selected if they were labeled as low quality by human annotators and correctly classified as low quality by the Random Forest model. Structural modifications were then applied using SHAP feature guidance derived from the Random Forest. Rows indicate transitions in categorical quality levels predicted by an **independent MLP** classifier (**0→1**, **1→2**, **0→2** = improvements; **Unchanged**; ↓ = degradations), while columns show the resulting impact on all evaluated quality dimensions. Percentages are computed over the subset of pages initially rated as low quality in the target dimension ($n$ denotes the number of such pages).

| Target | Change Type | Overall Quality | Relevance | Professionalism | Design | Authenticity |
|---|---|---|---|---|---|---|
| **Overall Quality** | 0→1 | **53.26** | 49.73 | 19.84 | 11.41 | 6.25 |
| | 1→2 | – | 3.53 | **10.87** | 1.09 | 0.00 |
| | 0→2 | **25.27** | 2.72 | 4.08 | 0.27 | 0.00 |
| ($n$=368) | Unchanged | **21.47** | 42.39 | 64.95 | 77.72 | 93.48 |
| | ↓ | **0.00** | 1.63 | 0.27 | 9.51 | 0.00 |
| **Relevance** | 0→1 | 46.94 | **77.55** | 0.00 | 0.68 | 0.00 |
| | 1→2 | 0.00 | – | **13.61** | 1.36 | 0.00 |
| | 0→2 | **22.45** | 6.12 | 0.00 | 0.68 | 0.00 |
| ($n$=147) | Unchanged | 30.61 | **16.33** | 85.71 | 91.16 | 100.00 |
| | ↓ | 0.00 | **0.00** | 0.68 | 6.12 | 0.00 |
| **Professionalism** | 0→1 | 57.58 | 0.00 | **72.73** | 6.06 | 0.00 |
| | 1→2 | 0.00 | **6.06** | – | 0.00 | 0.00 |
| | 0→2 | 12.12 | 0.00 | **15.15** | 0.00 | 0.00 |
| ($n$=33) | Unchanged | 30.30 | 93.94 | **12.12** | 78.79 | 100.00 |
| | ↓ | 0.00 | 0.00 | **0.00** | 15.15 | 0.00 |
| **Design** | 0→1 | 13.98 | 1.53 | 1.72 | **40.23** | 1.92 |
| | 1→2 | **9.20** | 0.77 | 5.94 | – | 0.00 |
| | 0→2 | **2.49** | 0.00 | 0.38 | 1.15 | 0.00 |
| ($n$=522) | Unchanged | 68.39 | 97.51 | 91.95 | **58.62** | 98.08 |
| | ↓ | 5.94 | 0.19 | 0.00 | **0.00** | 0.00 |
| **Authenticity** | 0→1 | 26.67 | 0.00 | 0.00 | 23.33 | **100.00** |
| | 1→2 | 10.00 | 6.67 | 13.33 | 0.00 | – |
| | 0→2 | **16.67** | 0.00 | 0.00 | 0.00 | 0.00 |
| ($n$=30) | Unchanged | 36.67 | 93.33 | 86.67 | 76.67 | **0.00** |
| | ↓ | 10.00 | 0.00 | 0.00 | 6.67 | **0.00** |

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 19b).

### 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 feature: `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 19c).

### 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 19d).

## 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`

Table 12: Cross-dimensional quality transition matrix after SHAP-guided HTML optimization. For each primary dimension (left), pages were selected if they were labeled as low quality by human annotators and correctly classified as low quality by the Random Forest model. Structural modifications were then applied using SHAP feature guidance derived from the Random Forest. Rows indicate transitions in ordinal quality levels predicted by an **independent ordinal** classification model (**0→1**, **1→2**, **0→2** = improvements; **Unchanged**; ↓ = degradations), while columns show the resulting impact on all evaluated quality dimensions. Percentages are computed over the subset of pages initially rated as low quality in the target dimension ($n$ denotes the number of such pages).

| Target | Change Type | Overall Quality | Relevance | Professionalism | Design | Authenticity |
|---|---|---|---|---|---|---|
| **Overall Quality** ($n$=368) | 0→1 | **91.30** | 54.89 | 23.10 | 14.40 | 6.52 |
| | 1→2 | – | 0.54 | **2.99** | 0.00 | 0.00 |
| | 0→2 | **1.90** | 0.00 | 1.63 | 0.00 | 0.00 |
| | Unchanged | **6.79** | 43.48 | 72.28 | 81.79 | 93.48 |
| | ↓ | **0.00** | 1.09 | 0.00 | 3.80 | 0.00 |
| **Relevance** ($n$=147) | 0→1 | 78.91 | **88.44** | 0.00 | 1.36 | 0.00 |
| | 1→2 | 0.00 | – | **5.44** | 0.00 | 0.00 |
| | 0→2 | **2.04** | 0.00 | 0.00 | 0.00 | 0.00 |
| | Unchanged | 19.05 | **11.56** | 94.56 | 94.56 | 100.00 |
| | ↓ | 0.00 | **0.00** | 0.00 | 4.08 | 0.00 |
| **Professionalism** ($n$=33) | 0→1 | 93.94 | 0.00 | **93.94** | 6.06 | 0.00 |
| | 1→2 | 0.00 | **3.03** | – | 0.00 | 0.00 |
| | 0→2 | 3.03 | 0.00 | **6.06** | 0.00 | 0.00 |
| | Unchanged | 3.03 | 96.97 | **0.00** | 84.85 | 100.00 |
| | ↓ | 0.00 | 0.00 | **0.00** | 9.09 | 0.00 |
| **Design** ($n$=522) | 0→1 | 18.77 | 1.92 | 2.11 | **60.73** | 1.72 |
| | 1→2 | 0.57 | 0.19 | **1.15** | – | 0.00 |
| | 0→2 | 0.00 | 0.00 | 0.00 | **0.00** | 0.00 |
| | Unchanged | 79.50 | 97.89 | 96.74 | **39.27** | 98.28 |
| | ↓ | 1.15 | 0.00 | 0.00 | **0.00** | 0.00 |
| **Authenticity** ($n$=30) | 0→1 | 56.67 | 0.00 | 0.00 | 23.33 | **96.67** |
| | 1→2 | 0.00 | 0.00 | **3.33** | 0.00 | – |
| | 0→2 | 0.00 | 0.00 | 0.00 | 0.00 | **0.00** |
| | Unchanged | 40.00 | 100.00 | 96.67 | 76.67 | **3.33** |
| | ↓ | 3.33 | 0.00 | 0.00 | 6.67 | **0.00** |

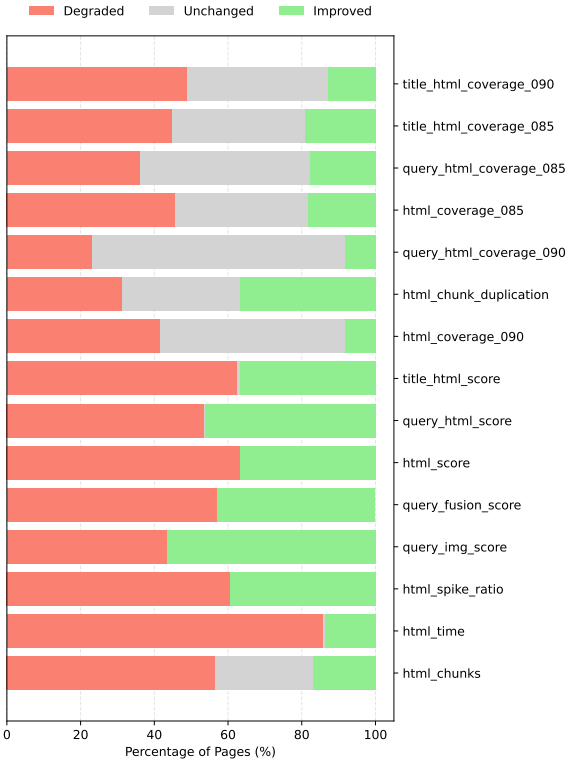
Table 13: Delta-based retrievability comparison across embedders. Positive values of ΔRank indicate improvement after HTML optimization, while negative values indicate degraded retrievability. For ΔRecall@10 and ΔMRR, the opposite holds.

| Embedder | Dimension | ΔRank | ΔRecall@10 | ΔMRR |
|---|---|---|---|---|
| `BAAI/bge-large-zh` | Overall Quality | -0.26 | 0.05 | 0.05 |
| | Relevance | -0.25 | 0.05 | 0.04 |
| | Professionalism | -0.31 | 0.04 | -0.01 |
| | Design | -0.39 | 0.06 | 0.05 |
| | Authenticity | -0.08 | 0.01 | -0.01 |
| `intfloat/multilingual-e5-large` | Overall Quality | -0.14 | 0.05 | 0.05 |
| | Relevance | -0.06 | 0.05 | 0.04 |
| | Professionalism | -0.10 | 0.04 | -0.01 |
| | Design | 0.02 | 0.06 | 0.05 |
| | authScore | 0.04 | 0.01 | -0.01 |
| `moka-ai/m3e-base` (used in our study) | Overall Quality | -0.28 | 0.02 | 0.04 |
| | Relevance | -0.44 | 0.003 | 0.05 |
| | Professionalism | -0.29 | 0.008 | 0.01 |
| | Design | -0.24 | 0.02 | 0.03 |
| | authScore | -0.04 | 0.00 | 0.002 |

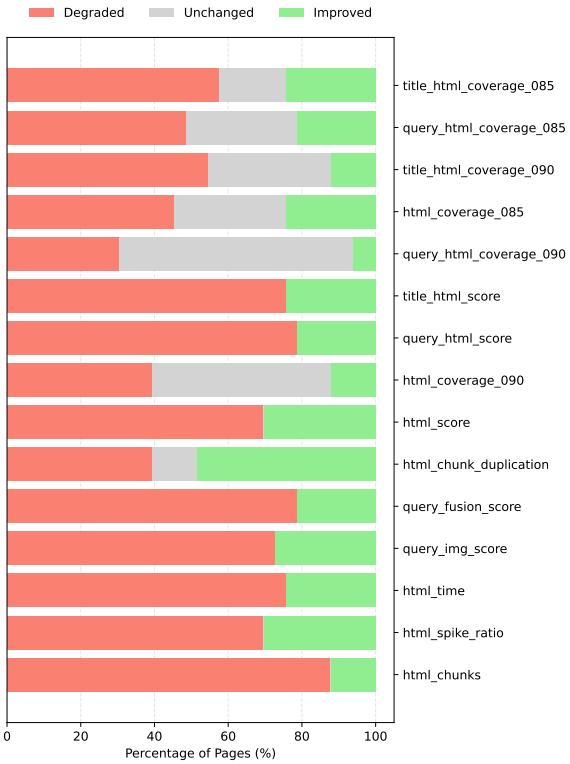
(retrieval-tuned)

- `moka-ai/m3e-base` (our main embedder)

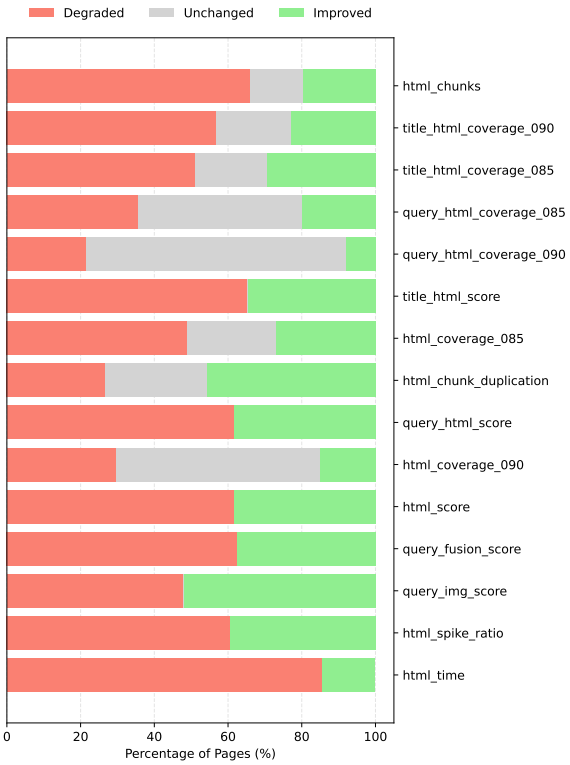
For each model, we build two FAISS indexes—



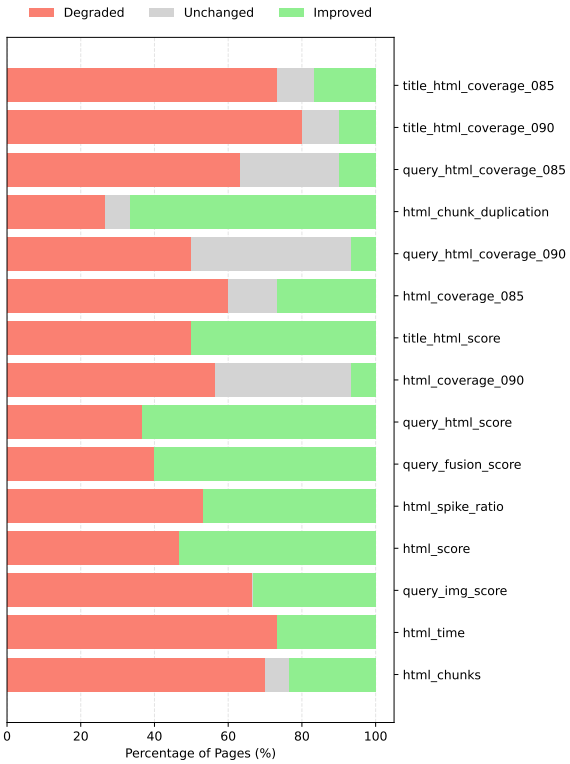

(a) **releScore:** Surface-level features occasionally improve, but semantic similarity degrades for the majority of pages (53–62%).




(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 19: 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.

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. Table 13 provides the aggregated deltas per dimension. Across all three embedders and four of the five WEBQUALITY dimensions, the optimized pages become *less* retrievable:

- ranks worsen ($\Delta$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.