International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Ai-Powered Website Seo Analysis And Intelligent Optimization

Preethi M.S., Moulya B, Ms Shivani Shikhka
3rd Year BCA, BMS College of Commerce and Management
Corresponding author: preethims2007@gmail.com
Co-author: moulya.hemavathi@gmail.com

Abstract - Search Engine Optimization (SEO) affects how easily websites are discovered, accessed, and used, yet conventional auditing typically requires several separate tools and expert interpretation, which limits its usefulness for non-technical users. This paper presents the design of a web-based system that combines automated on-page data extraction with Generative AI interpretation to produce structured website audits. Given a URL, the system retrieves the page, parses its HTML, and extracts meta tags, headings, links, images, word count, and related on-page attributes. Deterministic rule-based checks flag measurable problems, and the extracted data are then passed to a Generative AI model that groups findings into SEO, performance, accessibility, and best-practice categories, derives keyword-related insights, and drafts prioritized, plain-language recommendations. An overall assessment, issue list, and recommendations are consolidated into a single report. The contribution is not a new extraction or language-model technique but a documented integration of established components into one workflow that turns raw page data into understandable guidance. The paper describes the architecture, processing pipeline, implemented modules, and limitations, including model inaccuracy, dynamic content, and scoring subjectivity, and it specifies a test-case plan for future evaluation instead of reporting measured results. The approach is relevant to students, developers, website owners, and digital marketers who need an accessible starting point for improving website quality.

Keywords: Artificial Intelligence, Search Engine Optimization, Generative AI, Web Analysis, SEO Automation

1. INTRODUCTION

A website that cannot be found, read, or used gains little from its content. Search Engine Optimization (SEO) is the practice of structuring pages so that search engines can crawl, interpret, and present them, and so that visitors can use them effectively [1]. Search engines have long relied on page content and link structure to organize the web [2], which is why on-page elements such as titles, meta descriptions, headings, internal links, and image attributes remain central to SEO practice.

Auditing these elements is routinely done with browser developer tools, crawlers, and automated tools such as Lighthouse [3]. Such tools report many measurable findings, but the reader must still decide which findings matter, why they matter, and what to change. For students, small-site owners, and early-career developers, this interpretation step is often the main barrier, and it usually requires moving between several tools and reports.

Generative AI models, built on the Transformer architecture [6] and scaled as large language models [7], can read structured text and produce explanations in natural language. This suggests a division of labour: deterministic code extracts and measures page attributes reliably, while a language model organizes the findings and phrases recommendations in accessible

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

terms. The same models can, however, produce unsupported statements [9], so their output must be constrained by the extracted data and treated as advice to be verified.

Research problem. How can automated website data extraction and Generative AI be combined in one workflow so that a non-expert receives a structured, understandable, and data-grounded SEO audit from a single URL?

Objectives. The objectives of the study are as follows:

- To design a pipeline that extracts on-page SEO attributes from a URL.
- To combine rule-based checks with AI-based interpretation across SEO, performance, accessibility, and best-practice categories.
- To produce a consolidated report containing an overall assessment, issues, keyword-related insights, and recommendations.
- To define an evaluation plan and state the limitations of the approach honestly.

Contributions. The paper contributes (1) a documented architecture that integrates extraction, rule-based checks, and Generative AI interpretation in a single audit workflow; (2) a report structure that organizes findings by quality dimension with prioritized, plain-language recommendations; (3) an explicit separation of implemented features, proposed extensions, and future research; and (4) a test-case and evaluation plan. No measured performance results are claimed.

2. RELATED WORK AND LITERATURE REVIEW

2.1 SEO guidance and auditing tools. Search engine operators publish official guidance on crawlable structure, descriptive titles and snippets, heading use, link text, and image alt text [1]. This guidance is authoritative but descriptive, so a reader must compare a site against it manually. Lighthouse automates audits covering performance, accessibility, best practices, and SEO, and reports them by category with pointers to documentation [3]. Its audits are rule-based and do not produce tailored natural-language explanations of a specific site.

2.2 Crawling and on-page analysis. The architecture of early web search engines established crawling, parsing, and link-based analysis as the basis of web indexing [2]. On-page auditing applies the same idea at a smaller scale: fetch a page, parse its markup, and inspect the elements that search engines and users rely on. The proposed system uses this established approach as its data-collection stage and does not present it as new.

2.3 Accessibility and performance standards. The Web Content Accessibility Guidelines define testable success criteria, including text alternatives for images and meaningful heading and link structure [4]. Web performance is commonly discussed through user-centred metrics such as the Core Web Vitals [5]. Some accessibility criteria can be partly checked from static HTML, whereas most performance metrics need a real browser load, which limits what a lightweight analyzer can responsibly claim (Section 8).

2.4 Generative AI for analysis and recommendation. Large language models can adapt to tasks from instructions and examples given in a prompt [7]. Grounding generation in supplied or retrieved information is a recognised way to improve factual reliability [8], while surveys of hallucination document that generated text may be fluent yet unsupported by its source [9].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Separately, work on generative engine optimization studies how content can be adapted to be surfaced by generative search systems [10]. That work addresses visibility within AI-generated answers rather than auditing on-page quality, so it is related but distinct.

2.5 Comparison and gaps. Table 1 compares the approaches above with the proposed system. The comparison is limited to what the cited sources state.

Table 1. Comparison of relevant approaches and the proposed system

Approach	Primary focus	Automated extraction	Multi-category audit	AI-generated interpretation
Official SEO guidance [1]	Best-practice description	No (manual comparison)	Partly (SEO topics)	No
Lighthouse [3]	Automated rule-based audits	Yes	Yes (four categories)	No
Generative engine optimization [10]	Content visibility in generative engines	Not an audit tool	No	Studies generative systems
Proposed system	Consolidated, explained audit from a URL	Yes (HTML parsing)	Yes (four categories)	Yes (Generative AI)

The gap addressed here is practical rather than theoretical. Existing audit tools measure well but leave interpretation to the user, while general-purpose language models can explain but are not tied to a specific page unless a pipeline supplies its extracted data. Peer-reviewed studies that directly evaluate AI-generated SEO audit recommendations were not identified among the sources cited here, which is a limitation of this review.

3. RESEARCH GAP AND NOVELTY

Established techniques. HTML retrieval and parsing, rule-based on-page checks, standard accessibility and performance criteria [3]–[5], and the use of large language models for text generation [7] are established and are not claimed as new.

Proposed contribution. The contribution is the integration of these elements into one workflow in which (a) extraction produces a structured representation of a page, (b) deterministic checks mark measurable issues, (c) a Generative AI model is instructed to interpret only that structured data and to organize findings by category with prioritized, plain-language recommendations, and (d) results are consolidated into a single report with an overall assessment.

Usability contribution. The intended value is reduced tool-switching and reduced interpretation effort for non-specialist users. This is a design hypothesis that must be tested through the evaluation plan in Section 6. The work does not claim that AI-powered SEO analysis is unprecedented, does not predict search rankings, and does not track live keyword positions.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

4. PROPOSED METHODOLOGY

4.1 System architecture. The system follows a client-server design, shown in Figure 1. A web client accepts a URL and displays reports. A back-end service validates input, retrieves and parses the page, runs rule-based checks, calls a Generative AI model, assembles the report, and stores it for later viewing.

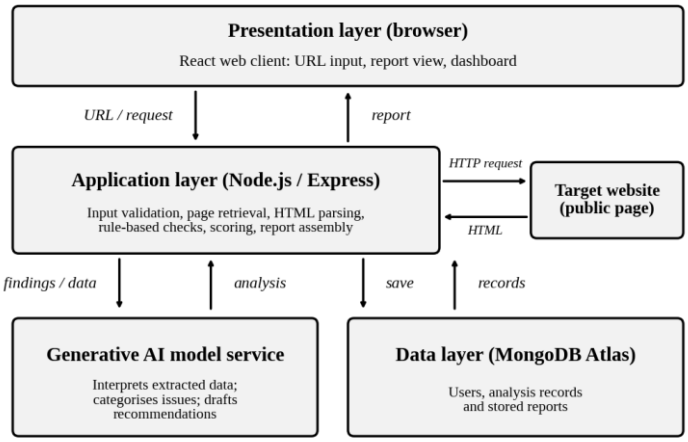


Figure 1. Proposed system architecture

4.2 Extraction. From the retrieved document the system extracts the page title, meta description and other meta tags, the heading hierarchy (H1–H6), internal and external links, images with their alt-text status, visible-text word count, and other supported on-page attributes. The output is a structured record per page.

4.3 Rule-based checks. Deterministic rules flag conditions that can be verified directly from the record, for example a missing title, a missing or repeated H1, images without alt text, or very short body text. Fixed, documented thresholds make these results reproducible.

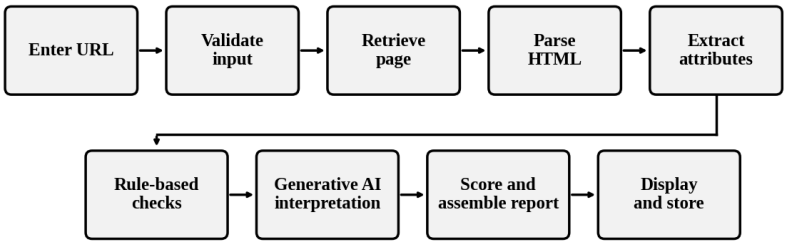
4.4 Generative AI analysis. The structured record and the rule-based findings are supplied to a Generative AI model with instructions to group issues into SEO, performance, accessibility, and best-practice categories, to give keyword-related observations derived from the page text, and to propose prioritized recommendations that refer only to the supplied data. Output is requested in a fixed structure so that it can be parsed and displayed consistently. Constraining the model to extracted data is intended to reduce unsupported statements [8], [9] but does not eliminate them.

4.5 Keyword-related analysis. Term frequency and keyword density are computed from the visible text, and the model comments on prominent terms and their placement in titles and headings. The system reports observations about the analysed page only. It does not estimate search volume, competitor position, or rankings.

4.6 Scoring and reporting. An overall assessment and per-category scores summarize the findings. The scores are heuristic indicators defined by the system's own rules and weights, not standardized measures, and they should not be read as predictions of search performance. The workflow is shown in Figure 2.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026



An error at any stage stops the workflow and returns a readable message to the user.

Figure 2. Website SEO analysis workflow

5. SYSTEM DESIGN AND IMPLEMENTATION

5.1 Technologies. The system is a web application with a React-based front end, a Node.js and Express back end that exposes a REST API, MongoDB Atlas for data storage, and web scraping with HTML parsing for data collection. Where a particular behaviour depends on configuration, it is described as such. Table 2 lists the functional modules and their responsibilities.

Table 2. Functional modules and their responsibilities

Module	Responsibility
User authentication	Registers and signs in users so that reports are linked to accounts.
SEO analyzer	Retrieves the page, parses the HTML, extracts elements and detects rule-based issues.
AI analysis	Interprets extracted data, categorizes issues and drafts recommendations.
SEO report	Presents scores, issues, keyword insights and recommendations.
Dashboard	Lets users view their analyses and previous reports.
Database management	Stores user accounts and analysis records.

5.2 Input validation and error handling. A submitted URL is checked for valid format and a supported scheme before any request is made. Retrieval failures, timeouts, non-HTML responses, and malformed AI output should produce an explanatory message, or a report based on the rule-based findings alone, rather than a silent failure.

5.3 Implementation status. The modules in Table 2 are the implemented scope of the current build. Browser-rendered analysis, multi-page crawling, measured performance data, and report export are proposed extensions that are not presented as completed features. Section 8 lists further research directions.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

6. RESULTS AND DISCUSSION

No experiments have been completed, and no measured results, datasets, or user feedback are reported in this paper. This section instead defines how the approach should be evaluated. The expected outcomes in Table 3 are design specifications and are not observations.

Table 3. Planned test cases and expected outcomes (expected, not measured)

ID	Test case	Expected outcome	Measured result
T1	Well-formed page with title, meta description, single H1 and alt text	Few or no rule-based issues; report completes	Not yet measured
T2	Page missing title and meta description	Both flagged by rules; recommendations refer to them	Not yet measured
T3	Page with images lacking alt text	Images counted and flagged under accessibility	Not yet measured
T4	Invalid or unreachable URL	Readable error message; no report generated	Not yet measured
T5	Client-side-rendered page with little static HTML	Limitation noted; sparse extraction reported honestly	Not yet measured
T6	Same page analysed repeatedly	Rule-based results identical; AI wording may vary	Not yet measured
T7	Malformed AI response	Fallback to rule-based findings	Not yet measured

Evaluation plan. The plan has five parts. (a) Extraction correctness: compare extracted counts with manual inspection on a small, documented set of pages. (b) Faithfulness of recommendations: check each recommendation against the extracted data and record unsupported statements [9]. (c) Consistency: repeat analyses and compare rule-based and AI outputs. (d) Agreement with an established tool: compare issue categories with Lighthouse output [3] and report both agreement and disagreement. (e) Usability: run a small study with students and site owners measuring understanding and task completion. Sample sizes, pages, and criteria must be fixed before these evaluations are run.

Discussion. Without measurement, claims about usefulness remain hypotheses. The design makes two properties inspectable: rule-based findings are reproducible by construction, and each AI recommendation can be traced to supplied data. Interpretability therefore depends on how faithfully the model follows its instructions, which is what evaluation (b) tests.

7. NOVELTY AND PRACTICAL APPLICATIONS

The integrated workflow is intended for four groups. Students can see how page attributes relate to quality criteria in plain language. Developers can obtain a first-pass review before using specialist tools. Website owners receive prioritized actions without interpreting raw reports. Digital marketers can use the output as a starting checklist for on-page review. For accessibility, the system can surface issues detectable in static markup, such as missing

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

alternative text [4], but it does not certify conformance, which requires manual and assistive-technology testing. Usability benefits are expected rather than demonstrated until the study in Section 6 is performed.

8. LIMITATIONS AND FUTURE WORK

Limitations.

- AI inaccuracy: generated text may be incomplete, generic, or unsupported [9], so recommendations should be verified.
- Crawling restrictions: sites may block automated requests, require authentication, or disallow crawling.
- Dynamic websites: pages built on the client side may expose little content to a static fetch.
- Rate limits: model and network calls are constrained by provider limits.
- Scoring subjectivity: weights and thresholds are design choices and have not been validated.
- Performance: HTML analysis cannot measure real loading behaviour such as Core Web Vitals [5], so performance findings are indicative only.
- Scope: analysis is per URL, and the system makes no ranking predictions.

Future work. Possible enhancements, kept separate from implemented features, are browser-rendered analysis for dynamic pages, integration of measured performance data [3], [5], site-wide crawling, scoring validated against expert audits, retrieval-grounded recommendations [8], and a study of how audits relate to visibility in generative search [10].

9. CONCLUSION

This paper presented the design of an AI-assisted website SEO auditing workflow that combines automated on-page extraction, rule-based checks, and Generative AI interpretation in one consolidated report. Its contribution lies in the integration and in making findings understandable to non-specialists, not in new extraction or language-model techniques. Because no evaluation has yet been completed, the benefits remain to be tested, and the evaluation plan and main limitations have been stated explicitly. Future work should measure extraction accuracy, recommendation faithfulness, consistency, and user comprehension before stronger claims are made.

REFERENCES

[1] Google, “SEO Starter Guide.” [Online]. Available: https://developers.google.com/search/docs/fundamentals/seo-starter-guide

[2] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” Comput. Netw. ISDN Syst., vol. 30, no. 1–7, pp. 107–117, 1998.

[3] Google Chrome Developers, “Lighthouse overview.” [Online]. Available: https://developer.chrome.com/docs/lighthouse/overview

[4] W3C, “Web Content Accessibility Guidelines (WCAG) 2.2,” W3C Recommendation, 2023. [Online]. Available: https://www.w3.org/TR/WCAG22/

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

[5] Google, “Web Vitals.” [Online]. Available: https://web.dev/articles/vitals

[6] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.

[7] T. B. Brown et al., “Language models are few-shot learners,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 1877–1901.

[8] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[9] Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Comput. Surv., vol. 55, no. 12, Art. 248, 2023.

[10] P. Aggarwal et al., “GEO: Generative engine optimization,” in Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2024.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)