An Intelligent Multi-Source Information Retrieval System for Hospital Chain Comparison using LLM-based Analysis

Apaar Raina, Aadish Jain, Krishna Agarwal

Abstract

This paper presents an intelligent information retrieval system that aggregates multi-source data from online reviews and news articles to generate comparative evaluations of hospital chains. Our approach employs dynamic web scraping using Selenium to collect reviews from platforms including JustDial and MouthShut, along with news headlines from Google News. The collected data is processed through six Large Language Models (DeepSeek Chat, GPT-4o Mini, Llama 3.1 70B, Llama 3.3 70B, Qwen 2.5 72B, and Claude 3.5 Haiku) to perform sentiment analysis and generate comprehensive comparative reports across multiple dimensions including patient care quality, infrastructure, staff behavior, and reputation. Comparative evaluation reveals significant differences in model performance, with DeepSeek Chat demonstrating superior critical analysis and patient safety focus. The modular architecture enables easy adaptation to other domains.

Keywords: Information Retrieval, Web Scraping, Large Language Models, Sentiment Analysis, Healthcare Analytics

1 Introduction

1.1 Background and Motivation

Healthcare decision-making significantly impacts patient outcomes, yet information about healthcare providers is fragmented across multiple online platforms. Patients face challenges in comprehensively evaluating hospital chains due to scattered reviews, varying rating systems, and lack of integrated analysis combining patient experiences with media coverage [1]. Traditional hospital evaluation methods focus on clinical outcomes while overlooking crucial patient experience dimensions.

Current approaches suffer from: (1) fragmented information across platforms, (2) time-intensive manual research, (3) subjective individual biases, (4) lack of temporal context, and (5) limited comparative analysis tools. The emergence of Large Language Models (LLMs) offers new opportunities for automated text analysis and synthesis [5].

1.2 Contributions

Our main contributions include: (1) a multi-source data aggregation framework using Selenium for dynamic scraping, (2) LLM-based comparative analysis using six models with varying capabilities, (3) multi-dimensional evaluation across patient care, infrastructure, staff behavior, and reputation, (4) comprehensive model performance comparison identifying critical thinking and patient safety assessment capabilities, (5) domain-agnostic modular architecture, and (6) qualitative validation methodology for LLM-generated reports.

2 Literature Review

2.1 Web Scraping and Data Collection

Mitchell introduced comprehensive web scraping approaches, emphasizing dynamic content extraction challenges [2]. Selenium WebDriver has emerged as powerful for JavaScript-rendered content. Glez-Peña et al. demonstrated that dynamic scraping is essential for modern web applications with client-side rendering [3].

2.2 Healthcare Sentiment Analysis

Greaves et al. analyzed patient feedback from online sources, demonstrating strong correlations between online ratings and official quality metrics [1]. Hao and Zhang proposed methods for mining healthcare reviews to extract patient experiences, emphasizing aspect-based sentiment analysis [4].

2.3 Large Language Models

Brown et al. introduced GPT-3, demonstrating few-shot learning capabilities for complex text analysis [5]. Touvron et al. presented Llama, an open-source LLM achieving competitive performance while being accessible to researchers [6]. Open-source models have democratized access to advanced NLP capabilities.

2.4 Multi-Source Aggregation

Zhang and Liu explored techniques for aggregating information from multiple sources, highlighting importance of source credibility and conflict resolution [7]. Dong et al. proposed knowledge fusion methods for integrating facts from sources with varying reliability [8].

2.5 Gap Analysis and Novelty

Existing research addresses individual components but gaps remain: (1) limited integration of reviews and news for entity comparison, (2) lack of LLM-based healthcare provider comparison systems, (3) insufficient multi-model

LLM approaches with critical safety evaluation, and (4) limited domain-agnostic frameworks.

Our project's novelty lies in combining real-time review scraping from multiple platforms with live news aggregation, and analyzing the collected data using six different LLMs to compare analytical capabilities, critical thinking, and patient safety focus. The system can be applied to any field by simply updating the input URLs. The system also provides clear reasoning using LLMs and uses Selenium to handle and extract data from websites that load content dynamically.

3 Objective

Primary Objective: Develop an intelligent IR system that automatically collects, processes, and synthesizes multi-source data to generate comprehensive hospital chain comparisons while evaluating LLM performance in critical healthcare analysis.

Secondary Objectives: (1) Implement robust Selenium-based web scraping for JustDial, MouthShut, and Google News, (2) integrate and compare six LLM models for analysis quality, (3) evaluate hospitals across patient care, infrastructure, staff behavior, and reputation, (4) generate human-readable comparative reports, (5) assess LLM output coherence, accuracy, and critical thinking capabilities, and (6) validate system modularity for domain adaptation.

4 Proposed Model

4.1 System Architecture

The system comprises 2 pipeline components:

Web Scraping Module: Dynamic Selenium-based scraper with multi-platform support, configurable URLs, and error handling. This saves the reviews and the news into json files to be given as input to the LLM.

LLM Analysis Engine: Works with DeepSeek Chat, GPT-4o Mini, Llama 3.1 70B, Llama 3.3 70B, Qwen 2.5 72B, and Claude 3.5 Haiku, uses prompt design for comparing results, and includes a framework to evaluate the data from multiple angles.

4.2 Data Flow

The sequential flow: (1) Input hospital names and URLs, (2) Selenium extracts reviews and news and stores them as json files (3) Six LLMs process the files with identical analysis prompts (4) outputs are saved as txt files for comparative evaluation.

5 Methodology

5.1 Web Scraping Implementation

We employ Selenium WebDriver for JavaScript-rendered content handling. The scraper implements browser automation, waits for dynamic loading, scrolls for infinite-scroll content, and extracts fully-rendered data.

5.2 LLM Integration

The system integrates six language models—*DeepSeek Chat*, *GPT-4o Mini*, *Llama 3.1 70B*, *Llama 3.3 70B*, *Qwen 2.5 72B*, and *Claude 3.5 Haiku*—each contributing different strengths and exhibiting distinct analytical patterns. A structured prompt is used to guide the models to compare multiple hospitals across key parameters, including patient satisfaction, staff quality, hygiene, infrastructure, emergency response, treatment specialties, affordability, wait times, and recent news.

The prompt instructs the model to generate: (1) overall rankings, (2) parameter-wise comparisons with evidence from reviews, (3) specialization-based insights, (4) strengths and weaknesses for each hospital, (5) patient-focused recommendations, (6) improvement suggestions, and (7) critical warnings. The analysis is required to be detailed, data-grounded, and supported by direct quotes from reviews, ensuring a thorough multi-dimensional evaluation.

5.3 Multi-Factor Analysis

Evaluation dimensions: *Patient Care* (treatment effectiveness, wait times, communication), *Infrastructure* (cleanliness, equipment, emergency services), *Staff Behavior* (professionalism, responsiveness, empathy), and *Reputation* (media sentiment, awards, compliance).

5.4 Evaluation Methodology

Manual evaluation assesses: *Coherence* (logical flow, consistency), *Relevance* (alignment with reviews, citation accuracy), *Comprehensiveness* (factor coverage, balanced perspectives), *Critical Thinking* (ability to identify safety concerns, willingness to provide clear warnings), and *Cross-Model Comparison* (consensus findings, divergent interpretations).

6 Experimentation and Results

6.1 Dataset Description

We collected review and news data for seven major hospital chains in Delhi: BLK, Medanta, Fortis, Max, Apollo, Narayana, and AIIMS. The dataset spans January–November 2024 and combines multi-platform user reviews with real-time news articles for each hospital. Table 1 summarizes the final dataset statistics.

Table 1: Dataset Statistics (Reviews and News per Hospital)

Hospital	Reviews	News	Avg. Length
BLK	79	41	775.66
Medanta	102	27	1067.97
Fortis	114	27	251.69
Max	206	54	410.83
Apollo	102	37	435.60
Narayana	110	46	471.09
AIIMS	102	50	162.30

Across all hospitals, review sources included JustDial and MouthShut. Most hospitals contained a balanced mix, with MouthShut providing longer, more narrative reviews, while JustDial reviews were generally shorter and more compact.

6.2 LLM Model Comparison

Six LLM models were evaluated on their ability to analyze hospital review data and generate comprehensive comparative reports. Table 2 summarizes the quantitative performance metrics.

Table 2: LLM Model Performance Comparison

Model	Time (s)	Quality
DeepSeek Chat	52.02	9.5
Qwen 2.5 72B	91.20	8.1
GPT-4o Mini	39.19	8.0
Llama 3.3 70B	46.88	7.4
Llama 3.1 70B	16.97	6.0
Claude 3.5 Haiku	18.18	6.4

Note: Quality score (out of 10) represents composite evaluation of coherence, relevance, critical thinking, and patient safety focus.

6.2.1 DeepSeek Chat

DeepSeek provided the most comprehensive and reliable analysis. Key strengths included: (1) willingness to identify dangerous hospitals clearly, (2) every ranking supported by patient quotes, (3) proper identification of AIIMS as top choice (9/10) and Medanta as most problematic (4/10), and (4) balanced acknowledgment of both strengths and critical safety concerns.

The model demonstrated superior critical thinking by not hedging diplomatically when patient safety demanded directness. For example, it explicitly warned against Medanta with evidence of "fatal negligence," "arrogant doctors," and "experimental treatment approaches." This patient-safety-first approach distinguished it from other models.

6.2.2 Qwen 2.5 72B

Qwen provided detailed analysis with extensive patient quotes and thorough documentation. However, it showed: (1) longest processing time (91.20s), (2) tendency toward overly optimistic rankings, (3) hesitancy to strongly criticize hospitals despite evidence, and (4) truncated output suggesting technical limitations.

Quality scores were high for detail but lower for critical assessment, suggesting the model prioritizes comprehensive documentation over decisive evaluation.

6.2.3 GPT-4o Mini

GPT-4o Mini delivered balanced analysis with reasonable processing speed. Strengths included proper citation and acknowledgment of both positive and negative aspects. However, critical weaknesses emerged: (1) Max Hospital ranked too high (8.5/10) despite serious complaints, (2) AIIMS ranked lower than evidence warranted, and (3) insufficient criticism of hospitals with documented safety concerns.

The model appeared to favor diplomatic balance over critical patient safety assessment, potentially misleading patients about serious risks.

6.2.4 Llama 3.3 70B

Llama 3.3 produced clearly structured analysis with reasonable processing time. It correctly identified top performers like AIIMS but exhibited: (1) sparse justifications, (2) minimal use of supporting quotes, (3) formatting errors (Medanta score left blank), and (4) avoidance of specific criticism.

The analysis was functional but lacked depth necessary for informed medical decision-making.

6.2.5 Llama 3.1 70B

Llama 3.1 was the fastest model (16.97s) but sacrificed accuracy for speed. Critical problems included: (1) contradictory rankings (Max ranked #1 despite complaints), (2) Apollo ranked lowest (4.2/10) despite positive reviews, (3) recommendations contradicting its own rankings, and (4) inconsistent logical flow.

The model demonstrated that processing speed alone does not ensure analytical quality, particularly for critical healthcare evaluation.

6.2.6 Claude 3.5 Haiku

Claude Haiku provided fast processing but superficial analysis. Weaknesses included: (1) Apollo ranked first (7.5/10) without strong justification, (2) minimal supporting quotes, (3) generic recommendations, and (4) insufficient depth for comparing hospitals.

The analysis would not adequately inform real patient decisions, suggesting the lightweight model sacrificed analytical capability for speed.

6.3 Hospital Ranking Consensus

Analysis of all six models revealed consensus on certain hospitals while exposing divergence on others. Table 3 presents the consolidated rankings.

Table 3: Hospital Rankings Across Models (Average Scores)

Hospital	Avg Score	Range	Consensus
AIIMS	8.5	8.0–9.0	Strong
Narayana	7.2	6.5–7.5	Strong
Max	7.0	6.5–8.5	Weak
Apollo	6.6	4.2–7.5	Very Weak
Fortis	6.2	5.0–8.0	Weak
BLK	5.4	3.0–7.0	Moderate
Medanta	4.8	4.0–6.0	Moderate

Strong Consensus: All models agreed AIIMS provides the best overall care, combining experienced doctors, affordability, and patient-centered treatment. Narayana received consistent recognition for cardiac and pediatric specialization.

Weak Consensus: Max, Apollo, and Fortis showed high variance across models. Some models emphasized positive infrastructure while others focused on billing issues and ICU neglect. This suggests these hospitals have highly variable experiences depending on department and specific doctor.

Safety Concerns: DeepSeek and Qwen clearly identified Medanta as problematic, while other models were more diplomatic. BLK showed moderate consensus on quality concerns, particularly regarding surgical procedures and misdiagnosis.

6.4 Critical Thinking Assessment

Models differed significantly in their willingness to provide clear safety warnings:

High Critical Thinking: DeepSeek explicitly identified dangerous patterns at Medanta (patient deaths, negligence, arrogant doctors) and BLK (surgical errors, misdiagnosis). It provided actionable "avoid" recommendations with supporting evidence.

Moderate Critical Thinking: Qwen and GPT-4o Mini acknowledged problems but used softer language. They documented concerns without providing clear directional guidance to patients.

Low Critical Thinking: Llama models and Claude Haiku avoided strong criticism, even when evidence clearly indicated patient safety risks. Their analyses prioritized diplomatic balance over patient protection.

6.5 Specialty-Specific Recommendations

Consensus emerged across models for specialty care:

Emergency Care: AIIMS (all models agreed on efficient emergency response and experienced staff).

Cardiac Surgery: Narayana (unanimous recognition of cardiac specialization and successful outcomes).

Budget Patients: AIIMS (clear consensus on affordability without quality compromise).

Planned Surgery: Mixed recommendations (DeepSeek: Narayana/Apollo with verification; Others: Max Hospital).

ICU Care: Avoid Fortis and Max at certain locations (consistent criticism of ICU neglect across multiple models).

6.6 Discussion

The multi-model evaluation revealed that analytical approach matters more than processing speed for healthcare assessment. DeepSeek's superior performance stemmed from: (1) critical thinking without diplomatic hedging, (2) evidence-based claims with patient quotes, (3) patient safety prioritization, (4) comprehensive coverage, and (5) clear actionable guidance.

Faster models (Llama 3.1, Claude Haiku) sacrificed analytical depth, while verbose models (Qwen) sometimes missed decisive conclusions despite extensive documentation. The ideal model balances thoroughness with critical assessment focused on patient safety.

Key insight: For medical analysis, thoroughness and critical assessment matter more than processing speed or diplomatic language. When someone's health is at stake, AI models must prioritize accuracy and patient safety over institutional politeness.

7 Conclusions and Limitations

7.1 Conclusions

This work demonstrates that automated hospital-chain comparison is feasible when combining multi-source data aggregation with multi-LLM reasoning. Key conclusions include:

(1) **Model Performance Varies Significantly:** DeepSeek Chat provided superior critical analysis and patient safety focus, while faster models sacrificed analytical quality. Model selection critically impacts the reliability of healthcare assessments.

(2) **Critical Thinking Essential for Healthcare:** Models must be willing to clearly identify safety concerns rather than diplomatically hedging. Patient protection requires direct language about dangerous patterns, not balanced "both sides" presentations when evidence is clear.

(3) **Value of Multi-Model Validation:** Using six models revealed consensus areas (AIIMS, Narayana) and uncertainty areas (Max, Apollo, Fortis), helping identify which hospital evaluations are reliable versus requiring additional research.

(4) **Multi-Source Aggregation Improves Reliability:** Combining patient reviews with news articles provided holistic evaluation capturing individual experiences and organizational patterns.

(5) **Domain-Agnostic Architecture:** The modular design enables rapid adaptation to other comparison domains with minimal configuration changes.

(6) **Practical Patient Value:** Generated rankings, warnings, and specialty recommendations support informed healthcare decisions, replacing hours of manual research.

7.2 Limitations

(1) **Scraping Robustness:** Anti-bot mechanisms require retry logic and occasional manual intervention.

(2) **Lack of Objective Ground Truth:** Manual validation introduces evaluator bias. Future work should incorporate quantitative medical outcomes or expert annotations.

(3) **Temporal Skew:** Recent reviews may disproportionately affect trends. Temporal weighting would reduce recency bias.

(4) **LLM Hallucination Risk:** Models sometimes infer details not present in data. Stricter citation enforcement needed.

(5) **Computational Cost:** High-quality models like DeepSeek incur significant cost. Local models like Ollama reduce cost but sacrifice depth.

(6) **Manual Data Updates:** Automated scheduled scraping would ensure real-time evaluation alignment.

(7) **Review Authenticity:** No detection of fake or incentivized reviews implemented.

(8) **Limited Human Evaluation:** Large-scale expert validation would strengthen confidence in assessments.

8 References

References

[1] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. and Donaldson, L., 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research*, 15(11), p.e239.

[2] Mitchell, R., 2018. *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly Media, Inc.

[3] Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. and Fdez-Riverola, F., 2014. Web scraping technologies in an API world. *Briefings in Bioinformatics*, 15(5), pp.788-797.

[4] Hao, H. and Zhang, K., 2016. The voice of Chinese health consumers: a text mining approach to web-based physician reviews. *Journal of Medical Internet Research*, 18(5), p.e108.

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., et al., 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, pp.1877-1901.

[6] Touvron, H., Lavril, T., Izacard, G., Martinet, X., et al., 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[7] Zhang, Y. and Liu, B., 2014. Multi-source information fusion based on rough set theory: A review. *Information Fusion*, 17, pp.59-71.

[8] Dong, X.L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S. and Zhang, W., 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9), pp.938-949.