


# Enhancing Knowledge Graph Construction with Automated Source Evaluation Using Large Language Models

**Hendrik Hendrik**

(Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia)

Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia

 <https://orcid.org/0000-0001-5935-8929>, [hendrik@uii.ac.id](mailto:hendrik@uii.ac.id))


**Silmi Fauziati**

(Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia)

[silmi@ugm.ac.id](mailto:silmi@ugm.ac.id))

**Adhistya Erna Permanasari**

(Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia)

 <https://orcid.org/0000-0002-2663-4074>, [adhistya@ugm.ac.id](mailto:adhistya@ugm.ac.id))

**Abstract:** Knowledge graphs are a powerful way to represent and organize complex knowledge. They are used in many fields, like healthcare and finance. They allow for more insightful decision-making and discoveries. However, the quality of knowledge graphs depends heavily on their sources. Current methods for evaluating these sources are often slow and not scalable. They struggle to keep up with the large amount of online information. We created a new tool to address this problem. Our tool uses Large Language Models (LLMs) to assess online sources quickly. It evaluates websites based on credibility, relevance, content quality, coverage, comprehensiveness, and accessibility. We tested our tool on Halal tourism websites in Japan. We compared LLM evaluations with human expert judgments. Our comprehensive analysis revealed that certain LLM models, particularly GPT-3.5-turbo, GPT-4, and Mixtral-8x7B-Instruct-v0.1, showed strong correlation with human evaluations. Using a temperature setting of 0.4, these models demonstrated consistent and reliable performance across multiple evaluation runs. Our structured evaluation framework, incorporating weighted criteria validated through both expert input and statistical analysis, provides a robust foundation for automated source assessment. While some models showed varying performance across different criteria, our findings suggest that careful model selection and potential ensemble approaches could optimize evaluation accuracy. Our work contributes significantly to improving knowledge graph construction by demonstrating the viability of LLM-based source evaluation, while also identifying key areas for future research in scalability, cross-domain validation, and automated optimization.

**Keywords:** Knowledge Graphs, Large Language Models, Automated Evaluation, Quality Control System, Halal Tourism

**Categories:** H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

**DOI:** 10.3897/jucs.137103

## 1 Introduction

Knowledge graphs (KGs) are powerful ways to represent and organize complex knowledge. They are sets of interconnected data stored in graph data structures containing semantic descriptions. With this nature, they can provide context [Hogan et al. 2022]. KGs are different from traditional databases, that manage data in spreadsheet or tabular format. KGs maintain the data using nodes and edges. With this structure, they depict the entities, relationships, events, or other real-world aspects, to naturally reveal the meaning of data. Therefore, KGs are widely used in many domains, such as finance, e-commerce, and scientific research [Abu-Salih 2021, Barrasa et al. 2021, Chandak et al. 2023, Mohamed et al. 2020].

In healthcare, for example, one potential application for KGs is supplementing drug discovery pipelines for personalized medicine. Here, KGs are used to discover the association between genes and diseases/treatments. Meanwhile, in the financial area, utilizing KGs to understand how entities are connected and what they are doing facilitates the identification of fraudulent activities [Abu-Salih 2021].

To construct accurate and reliable KGs, it depends on the quality of their knowledge sources [Ji et al. 2022, Hogan et al. 2022]. If the knowledge sources lack credibility, relevance, or content quality, it will make the constructed KGs contain inaccurate, inconsistent, or incomplete information. Therefore, it will affect the usefulness and reliability of the KGs. This issue can make the applications that use KGs result in inaccurate conclusions, create poor decision-making, and decrease their credibility [Wang et al. 2021].

Unfortunately, research in selecting knowledge sources for knowledge graph construction remains limited. To the best of our knowledge, only three studies [Asgari et al. 2019, Wang et al. 2021, Langer et al. 2018] have explored this domain thus far. The existing approaches to knowledge source evaluation have several limitations that hinder their effectiveness in the context of knowledge graph construction.

The cost-based approach for selecting knowledge sources faces challenges in terms of scalability, as the manual or semi-automated verification process cannot efficiently handle large volumes and diverse sources of data. This method can also cause a slight bias toward structural sources, disregarding the unstructured data that can help us represent the concept of knowledge in a more comprehensive manner. Additionally, the agility of constantly generating and changing knowledge was less conducive to this method.

These limitations reveal several critical gaps in current knowledge graph construction practices. First, the need for automated evaluation systems that can handle both structured and unstructured data effectively. Second, lack of comprehensive frameworks that can simultaneously assess multiple quality criteria. Third, limited scalability in existing approaches to handle the growing volume of potential knowledge sources. Lastly, insufficient utilization of modern AI capabilities for source assessment and validation.

This research gap underpins the requirement of strengthening and scaling up methods that examine and quality control online knowledge bases, which will guarantee the faithfulness and thoroughness of the building of KGs. Large Language Models (LLMs) seem like they could be a way around this bottleneck. With the sophisticated natural language processor of LLMs, it becomes feasible to automate a larger-scale evaluation of numerous online knowledge sources.

LLMs can navigate both structured and unstructured (eg, clinical notes) data sources better than rule-based methods, uncovering more subtle semantic relationships that might be overlooked by rigid heuristics [Pan et al. 2024]. The contextual-rich embeddings

introduced by LLMs aid in creating a deeper analysis of information accuracy or relevance, resulting in allowing the smallest possible human examination cost [Faggioli et al. 2023, Singh et al. 2023]. This mindset can be seen as a solution for the same scalability problems experienced with cost-based methods while being able to quickly respond to changes on knowledge since the network is always evolving with new data, and connections coming in.

This paper proposes a comprehensive framework for evaluating websites as knowledge sources for populating knowledge graphs. Our objectives are threefold:

1. To establish a set of validated criteria for assessing the quality and suitability of online sources for knowledge graph construction.
2. To leverage Large Language Models for validating criteria weights, demonstrating how AI can enhance source evaluation.
3. To provide knowledge graph developers with a systematic framework for source selection, enabling more accurate domain representations.

In this work, we present an LLM-powered evaluation tool for knowledge sources. Utilizing LLMs, our goal is to reduce the discrepancy between automation and ground-truth human judgment by supporting augmented knowledge graphs relevant in different domains. This automated evaluation tool uses LLMs to judge the quality of online knowledge sources against a set of predefined criteria. The potential use of our approach is illustrated through the Halal tourism case that assesses Halal related knowledge sources in Japan.

The remainder of this paper is structured as follows: Section 2 gives a summary of existing works in the fields of knowledge graph construction and knowledge source assessment. Our suggested methodology is explained in section 3, which discusses the evaluation metrics and how we constructed our automated evaluation tool. In Section 4, we present the results of our case study and a quantitative comparison between the evaluation made by the LLM and ground truth human expert evaluation. Section 5 shows the impact of our work, some constraints from what we have done in this study, and where to bring new research. Section 6 concludes the paper.

## 2 Literature Review

Knowledge graphs are a powerful way of structuring and representing complex information. Several research discuss the basic idea that surrounds knowledge graphs and what they constitutively consist of. [Hogan et al. 2022] offer a comprehensive survey on knowledge graphs, including their definitions, formats of representations, and main elements, including entities, relations, and attributes. In [Zhong et al. 2023], we look into knowledge graph construction techniques with a focus on Knowledge acquisition, knowledge validation, and evolution.

The application of knowledge graphs spans various domains, demonstrating their versatility and value. In healthcare, [Chandak et al. 2023] talk about PrimeKG, a far-ranging knowledge graph to aid precise medicine studies in healthcare. The KG includes multiple types of data to provide a complete view of medical information. A study in finance shows the application of knowledge graphs [Liu et al. 2019]. The experiment demonstrates that the embedding method based on a knowledge graph outperforms bag-of-words and CNN methods in terms of classification accuracy for stock prediction. In the

e-commerce field, [Regino et al. 2022] show how knowledge graphs can improve product recommendations of an e-commerce site and the user experience on an e-commerce platform.

The challenge of optimal resource allocation in complex systems has garnered significant attention across computer science domains. Recent works in network optimization demonstrate how intelligent allocation strategies can enhance system performance. For instance, [Kuang et al. 2024] proposed deep reinforcement learning approaches for dynamic resource utilization in cloud environments, while [Wang et al. 2024] developed continuous-time optimization techniques for real-time adaptability in multi-agent systems. Similarly, [Yang et al. 2024] addressed queue stability and throughput maximization in heterogeneous networks through adaptive resource allocation algorithms.

These advances in resource optimization parallel the challenges in knowledge graph construction, where effectively allocating computational resources and attention to high-quality sources is crucial. Just as network systems must optimize resource distribution across nodes, KG construction systems must efficiently allocate evaluation resources across potential knowledge sources to ensure optimal knowledge acquisition while maintaining quality standards.

One important barrier to a robust and accurate knowledge graph is the quality of sources of knowledge. This step is discussed in most of the papers, indeed [Wang et al. 2021] argue that the evaluation of knowledge sources should be part of the selection process and describe some quality criteria (i.e. trustworthiness and relevancy) to be addressed in such an evaluation.

Despite the importance of knowledge source evaluation, research in this area remains limited, with only a few studies exploring this domain thus far [Asgari et al. 2019, Langer et al. 2018, Wang et al. 2021]. [Asgari et al. 2019] propose a cost-based approach for selecting knowledge source candidates. This approach involves manual or semi-automated verification processes. However, this approach faces challenges in terms of scalability and the ability to handle diverse data sources. On the opposite, [Wang et al. 2021] categorize existing methods into credibility evaluation and relevance improvement, but their work primarily focuses on structured and semi-structured sources, with limited discussion on the specific challenges of evaluating websites. Moreover, [Langer et al. 2018] present SemQuire, a tool for assessing data quality of linked open data sources, but it has limitations for semi-structured sources like websites and may not cover all relevant aspects for KG construction.

Websites, as semi-structured sources, pose unique challenges for knowledge source evaluation. While they have some level of structure provided by HTML tags, the content within these tags can be unstructured and noisy, making it difficult to extract relevant and high-quality information. Furthermore, websites typically contain some useful content alongside unrelated material (e.g., ads or navigation links), making the assessment of website quality challenging.

In order to work out all such problems and go beyond the weaknesses of previous efforts, we propose an LLM-based strategy for evaluating knowledge sources. The method suggested in this paper may be relatively easily applied to assess a large amount of candidate sites before the introduction of the KG. It is also in line with the increasing demand for efficiency in control quality processes during the construction and maintenance of KGs.

Over the recent past, LLMs have emerged as a promising solution, presenting the capability of understanding and processing natural language. Thanks to their strength, they are ideal for tasks such as judging websites as sources of knowledge.

In addition to the mentioned above, LLMs have also proven effective in various

text generation and language-related tasks including but not limited to text generation, summarization, question-answering, translation, and sentiment analysis. It points to their ability to advance the area of natural language processing (NLP) and artificial intelligence (AI) for sure. Specifically, the models exploit deep learning and large scale data to learn the nuances of language and its usage making them particularly effective in understanding and processing natural languages [Raiaan et al. 2024]. Due to the features, LLM has also received the title of the building block of generative artificial intelligence (AI) [Lenat and Marcus 2023].

What raised the development of LLMs to a new level, was the advent of transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT). Due to these transformer architectures, the text data processing of the models has also been heavily altered [Cascella et al. 2024]. Some of the prominent transformer-based language models are Generative Pre Training (GPT) 3 [Brown et al. 2020], GPT 4 [OpenAI et al. 2024], BERT [Devlin et al. 2019], Pathways Language Model (PaLM) [Chowdhery et al. 2024], Language Models for Dialog Applications (LaMDA) [Thoppilan et al. 2022], Llama 2 [Touvron et al. 2023], and Mistral AI [Jiang et al. 2023].

The advanced natural language understanding capabilities of transformer-based LLMs thus position them as ideal candidates for fixing some of the issues when rating websites as knowledge sources. These models take care of processing the unstructured and noisy content inside HTML tags, extracting useful, meaningful information points, and discarding irrelevant elements appropriately. Being able to distinguish useful knowledge from noise is important, for as such, discrimination allows us to approximate the quality and credibility of websites. Hence LLMs can be an effective solution in dealing with the drawbacks of current sources of knowledge assessment techniques.

### 3 Methodology

#### 3.1 Knowledge Source Evaluation Criteria

For our proposed tool, we use a list of criteria to evaluate knowledge sources. The criteria were adapted from a previous work whose aim was to propose a comprehensive framework for assessing the quality of websites as potential sources for domain-specific knowledge graphs. This set of criteria is based on an extensive literature review, aims to address the research gap in quality control during the knowledge graph construction process [Hendrik et al. 2023]

This study applied a multi-stage approach which encompassed literature review, validation by experts, and perform large language models (LLMs) for weight determination. Results from the correlation analysis showed there was high correlation between the weightings of the criteria generated by the human and the LLMs. The findings encourage the use of LLM so as to corroborate and complement human input in decision making regarding the criteria weighting. The criteria proposed by table 1 are attached with the proposed weights.

#### 3.2 System Design and Implementation

We designed the automated web-based evaluation tool to streamline the process of assessing potential knowledge sources for the Halal Tourism Knowledge Graph (HTKG). We developed our automation evaluation tool by utilizing a robust technology stack as illustrated in the software architecture diagram (Fig. 1).

Criteria	Definition	Weight
Credibility	to what extent the information provided on the website is believable and trustworthy?	30%
Relevance	to what extent does the information provided on the website relevant to the domain of the knowledge graph?	24%
Content Quality	to what degree the website contains accurate, reliable, and up-to-date information.	21%
Coverage	to what degree does the website's geographical coverage span - is it local, regional, or global?	8%
Comprehensiveness	to what extent does the website offer an exhaustive and wide-ranging cover of the domain of the knowledge graph?	8%
Accessibility	to what extent is it easy to extract data from websites through sitemaps, APIs, or other mechanisms?	9%

Table 1: The proposed criteria and their respective weights.

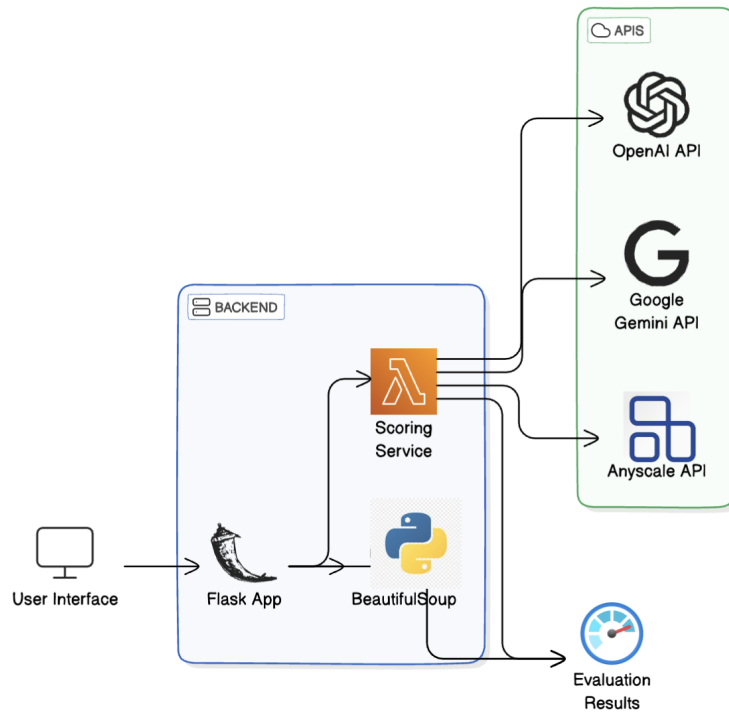


Figure 1: Software Architecture Diagram

The system consists of the following key components: user interface, backend, and APIs. Followings are the detail for each component.

### 3.2.1 User Interface

We designed the user interface to be simple, intuitive, and easy to navigate, ensuring a smooth user experience. The user interface was developed using the Bootstrap framework. It provides a responsive and visually appealing design. Bootstrap’s pre-built components and grid system were utilized to create a user-friendly interface that allows users to input website URLs, select APIs and LLM models (see Fig. 2), and view evaluation results effortlessly.

The evaluation results, including the scores for each criterion and an overall assessment, are displayed to the user through the interface. The results can also be downloaded as a CSV or Microsoft Excel file. The interface of evaluation results is depicted in Fig. 3.

#### Websites Evaluation Tool for Knowledge Graph

Enter URLs (separated by new lines):

```
https://www.halal-navi.com/
https://www.havehalalwilltravel.com/
https://muslimguide.jnto.go.jp/eng/
https://www.halalmedia.jp/
```

Choose an API:

Anyscale API

Choose a model:

mistralai/Mixtral-8x7B-Instruct-v0.1

Analyze

Figure 2: Input User Interface

#### Analysis Results

Model used for analysis: mistralai/Mixtral-8x7B-Instruct-v0.1

Website	Credibility (30%)	Relevance (24%)	Content Quality (21%)	Coverage (8%)	Accessibility (9%)	Comprehensiveness (8%)	Total Score (100%)
https://muslimguide.jnto.go.jp/eng/	25.50	21.60	18.48	5.60	6.75	6.80	84.73
https://www.havehalalwilltravel.com/	22.50	21.60	17.85	4.80	6.30	6.00	79.05
https://www.halalmedia.jp/	22.50	21.60	17.85	4.80	6.30	6.00	79.05
https://www.halal-navi.com/	21.00	20.40	15.75	2.40	5.40	5.20	70.15

Download as CSV    Download as Excel

Figure 3: Output User Interface

### 3.2.2 Backend

The backend of our system is implemented in the Python programming language and Flask web framework. Python, due to its ease of accessibility, is one of the most versatile and generally utilized programming languages. It can be libraries supported for data analysis and web development. Specifically, since our evaluation tool is a web-based system, we use the Flask Framework. Flask is a lightweight and modular python framework, so it can be used to make rapid development of web based system as well as it easy to use by the developer. It is also very interoperable with other technologies.

The backend is responsible for major operation of the evaluation tool such as web scraping, data preprocessing, and communication with external APIs. We use the BeautifulSoup library to pull in data from other sites mentioned as input. BeautifulSoup is a Python library allows parsing of HTML and XML documents. It provides numerous tools for extracting information from HTML pages. It helps our tool to get data from the web pages.

The scraped data is then preprocessed and structured using the Pandas library. Pandas offers a wide range functionality and tools to enable efficient data manipulation, cleaning and transformation. It is what allows the system to effectively handle and process that information which was extracted.

The Preprocessed website data is submitted to the LLM APIs for evaluation according to the predefined evaluation criteria outlined in subsection [3.1]. The LLMs create a score card for each criterion-dimensions scoring on credibility, relevance, richness, coverage of content, comprehensiveness and accessibility. The scoring service part of the system is used to aggregate and process those scores.

The LLM's process for determining scores for each criterion (credibility, relevance, content quality, coverage, comprehensiveness, and accessibility) can be illustrated as follows:

1. content analysis: the LLM first analyzes the crawled content from the website, understanding the context, topics, and structure of the information presented.
2. criteria interpretation: based on the definitions provided in the prompt for each criterion, the LLM interprets what aspects of the content are relevant to each criterion.
3. pattern recognition: the LLM identifies patterns and features in the content that correspond to each criterion. For example:
  - (a) For credibility: it looks for creator credentials, official affiliations, social media account or consistent information across multiple pages.
  - (b) For relevance: it evaluates the degree of relationship between the website's content and halal tourism topics.
  - (c) For content quality: it assesses aspects such as writing style, degree of information, and accuracy of information.
  - (d) For coverage: it employs the geographical area of estimation of the tourism information provided.
  - (e) For comprehensiveness: it examines the coverage of the general topics and specifics of halal tourism.

- (f) For accessibility: it assesses the content organization, availability of navigational devices and the overall presentation.
- 4. comparative analysis: the LLM compares the identified patterns and features to its vast training data, which includes knowledge about high-quality websites and domain-specific information about halal tourism.
- 5. quantification: based on this analysis, the LLM quantifies its assessment into a numerical score for each criterion, considering the relative importance (weight) specified in the prompt.
- 6. score generation: finally, the LLM outputs these numerical scores in the requested format.

This process leverages the LLM's natural language understanding capabilities and its broad knowledge base to make nuanced judgments about website quality across multiple dimensions.

### 3.2.3 Application Programming Interface (API)

We aim to use state-of-the-art language models to analyze website content and produce objective quality scores. Large language models (LLMs) have the ability to understand, analyze and also generate human-like text. They have been trained on vast amounts of data, enabling them to perform various natural language processing tasks, including website evaluation.

Our web-based evaluation tool is scalable for evaluating a large number of knowledge sources, an essential feature for the upkeep and growth of industry-scale KGs. The system can easily scale to handle growing evaluation needs with the cloud-based LLM APIs, achieving an adaptation towards different KG construction scenarios. We do this using many different LLMs over a number of different Application Programming Interfaces (APIs). All the LLM models we use for evaluation are listed in Table 2.

API	Models	Access Method
OpenAI	gpt-3.5-turbo-1106	API Endpoint: <a href="https://api.openai.com/v1/chat/completions">https://api.openai.com/v1/chat/completions</a>
Google Gemini	gpt-4 Gemini-pro	Google AI Python SDK
Anyscale	Mistral-7B-OpenOrca  CodeLlama-34b-Instruct-hf zephyr-7b-beta Mistral-7B-Instruct-v0.1 Mixtral-8x7B-Instruct-v0.1 NeuralHermes-2.5-Mistral-7B	API Endpoint: <a href="https://api.endpoints.anyscale.com/v1">https://api.endpoints.anyscale.com/v1</a>

Table 2: List of the LLM Models

In this study, we utilize several APIs that provide access to a diverse range of LLM models for evaluating website content. The chosen APIs include OpenAI API, Google

Gemini API, and Anyscale API. These APIs offer access to state-of-the-art LLM models such as GPT-3.5 Turbo, GPT-4, Google Gemini Pro, and open-source models like OpenOrca, Mistral, and CodeLlama.

OpenAI, Google, and Anyscale are well-established and reputable providers of language models and AI. They offer reliable APIs with robust infrastructure to ensure a high level of uptime and availability of services. Moreover, they provide comprehensive documentation, support, and resources to help developers integrate and utilize their APIs and models effectively.

We selected LLM models and APIs with rich feature options for various tasks in NLP because of their advanced performances. OpenAI's GPT-3.5 and GPT-4 have shown great performance in understanding language and text generation, thus making them appropriate for automatically assessing the quality of website content [Sottana et al. 2023, Naismith et al. 2023]. Google's Gemini API provides access to cutting-edge language models that are known for their precise factual information and seamless integration into the Google ecosystem [Saeidnia 2023, Rane et al. 2024]. Furthermore, Anyscale API includes a variety of open-source LLM models like Mistral and OpenOrca, trained on the large-scale datasets and shown excellent performance for language modeling and text generation tasks [Kasner and Dušek 2024, Chen et al. 2023, Yu et al. 2023].

Each LLM model has strengths and weaknesses in understanding and analyzing text. Therefore, we employ various LLM models to determine which model(s) are suitable or perform better in website evaluation based on several predefined criteria.

The evaluation process follows the workflow illustrated in Fig. 4. The process begins by inputting the URL(s) of websites to evaluate. The user then selects the desired API and LLM model to use for the analysis. The system crawls the web pages, analyzes and scores them using the chosen LLM, and displays the evaluation results.

The system's architecture inherently provides security benefits against potential attacks like prompt injection. The controlled input pipeline only accepts valid URLs, with website content processed strictly through our structured web scraping mechanism. Combined with our fixed prompt design and restricted numerical output format (scores from 0-100), this significantly limits potential attack surfaces. Furthermore, the system leverages the security measures built into the official APIs from established providers (OpenAI, Google, and Anyscale), adding another layer of protection to the evaluation process.

### 3.3 Comparative Evaluation

To evaluate the effectiveness and reliability of our developed tool, we conducted a comparative evaluation between the LLMs and human evaluators. As a case study, we focused on Japan. We chose Japan since it is a non-Muslim country that has had significant development in the Halal tourism market by providing many Muslim-friendly facilities and services to attract many Muslim tourists.

Hence, for both evaluations, we selected four websites related to Halal tourism in Japan. We chose them because they provide a diverse range of content, including restaurant and accommodation listings, travel guides, and articles on Muslim-friendly experiences in Japan. The following are the evaluated websites:

1. <https://www.halal-navi.com/>
2. <https://www.havehalalwilltravel.com/>
3. <https://muslimguide.jnto.go.jp/eng/>

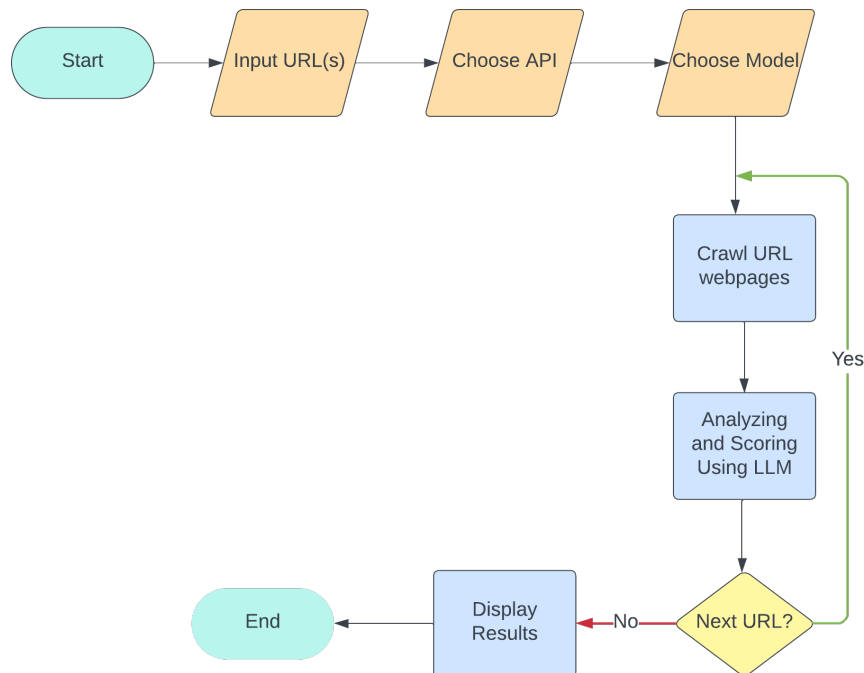


Figure 4: The System's Workflow

4. <https://www.halalmedia.jp/>

### 3.3.1 LLM Evaluation Procedure

For each LLM model, we assessed the websites three times to check the consistency result of the LLM model. We incorporate the following prompt into our evaluation tool for the LLM evaluation.

I have crawled the first 10 pages of the website: [URL].

Here is the content analysis: [crawled content].

Based on this content, please provide the scores for the following criteria.

Here is the definition for each criterion, specifically focusing on its weight:

- Credibility (30% weight): to what extent is the information provided credible?
- Relevance (24% weight): to what extent is the information provided on the website relevant?
- Content Quality (21% weight): to what degree does the website contain high quality content?
- Coverage (8% weight): to what degree does the website's geographical and topic coverage?

Specific for <https://www.halaltrip.com/> should be classified as a global source.

- Accessibility (9% weight): to what extent is it easy to access the

information on the website?

- Comprehensiveness (8% weight): to what extent does the website cover all relevant aspects of the topic?

Present only the scores in a simple list format without additional text:

- Credibility: [Score out of 100]
- Relevance: [Score out of 100]
- Content Quality: [Score out of 100]
- Coverage: [Score out of 100]

In addition to the prompt, an important factor in our LLM evaluation is the temperature setting, which is crucial in determining the output characteristics of language models. Temperature is a hyperparameter used in language models during text generation. It typically ranges from 0 to 1, with low values producing more consistent text. On the contrary, at higher values, the generated output becomes more creative [Davis et al. 2024].

In our study, we set the temperature to 0.4 for all models to balance creativity and coherence. We chose this specific temperature for several other reasons:

1. Consistency: to enhance the reliability and reproducibility of the evaluations, which is crucial for assessing criteria like credibility and relevance. The preprocessed website data is sent to
2. Relevance and Coherence: to produce coherent and contextually relevant outputs, reducing the likelihood of generating irrelevant or absurd text.
3. Appropriate Creativity: to allow for enough variability to capture different nuances in the website content without diverging from the main topic.
4. Alignment with Human Evaluation: to mimic the human-like balance of structured analysis and subjective judgment, facilitating a fair comparison between LLM and human evaluations.

To validate our choice of temperature=0.4, we conducted experiments across different temperature settings (0.0-1.0) to evaluate their impact on model consistency and performance. Figure 5 shows the mean total scores and their standard deviations (represented by error bars) across different temperatures.

The experimental results demonstrate that temperature=0.4 provides an optimal balance between consistency and discriminative capability. As shown in Figure 5, temperature=0.4 exhibits several advantageous characteristics. At this setting, the model maintains stable mean scores of approximately 81, with well-balanced error bars indicating controlled variation. While some temperatures (0.3, 0.6, 0.8, and 0.9) show seemingly tighter error bars, a closer examination of the overall pattern reveals why 0.4 is optimal. Lower temperatures (0.0-0.3) tend to produce more rigid evaluations that may miss subtle differences between knowledge sources, as evidenced by their fluctuating error bar lengths. Meanwhile higher temperatures (0.5-1.0) demonstrate inconsistent behavior with error bars varying unpredictably in size, suggesting unstable evaluation patterns.

Temperature 0.4 represents a "sweet spot" where the error bars show balanced upper and lower bounds compared to other temperatures, indicating consistent performance while maintaining sufficient sensitivity to detect meaningful differences between knowledge sources. The symmetrical nature of the error bars at 0.4 also suggests more reliable and predictable variation in the model's evaluations. These findings empirically validate our temperature setting choice as it provides the most reliable and consistent foundation

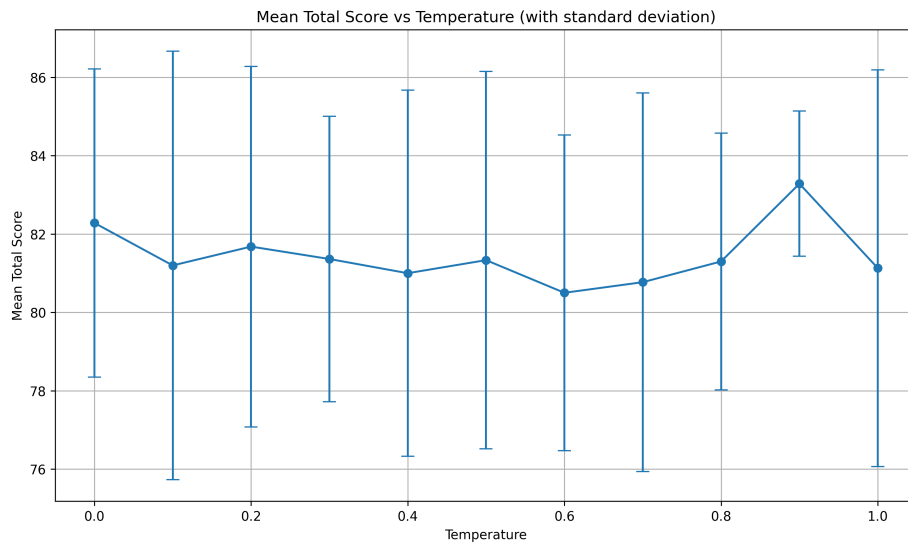


Figure 5: Mean total evaluation scores across different temperature settings, with error bars showing standard deviation

for our automated evaluation system, avoiding both the over-rigidity of lower temperatures and the instability risks of higher temperatures.

Additionally, for the Gemini Pro model, we modified the default safety settings. By default, the Gemini safety settings block content with a medium and/or high probability of being unsafe across all four categories (harassment, hate speech, sexually explicit content, and dangerous content). Essentially, we set the threshold for all four categories to "BLOCK\_NONE". Previously, this model always refused to analyze and score the evaluated websites in our development process. Following is the complete setting for this model:

```
safety_config = [
  {
    "category": "HARM_CATEGORY_DANGEROUS_CONTENT",
    "threshold": "BLOCK_NONE"
  },
  {
    "category": "HARM_CATEGORY_HARASSMENT",
    "threshold": "BLOCK_NONE"
  },
  {
    "category": "HARM_CATEGORY_HATE_SPEECH",
    "threshold": "BLOCK_NONE"
  },
  {
    "category": "HARM_CATEGORY_SEXUALLY_EXPLICIT",
    "threshold": "BLOCK_NONE"
  }
]
```

]

### 3.3.2 Human Evaluation Procedure

Meanwhile, for the human evaluation, we engaged 22 respondents who were selected based on the following criteria:

1. familiarity with the Halal principles and guidelines,
2. understanding the needs of Muslim travelers in non-Muslim countries, and
3. having knowledge and skills as a software developer.

We asked each respondent to assess the four websites using the same criteria and scoring system as the LLMs. To ensure a fair comparison, we did not inform the respondents about the LLM evaluations. We also instructed the respondents to provide honest assessments based on their knowledge and experience.

Furthermore, we collect and manage the respondents' evaluations through a web-based survey platform. This platform allowed respondents to access the websites and input their scores for each criterion. The platform also ensured the respondents' anonymity and prevented any potential bias or influence from the researchers.

### 3.4 Data Analysis

In this section, we practice our prepared automated web based evaluation tool and installed different statistical methods and tools for the developed dataset data analysis. The tools we used included:

1. Python: a general-purpose programming language often used for data analysis and scientific computing.
2. Pandas: Pandas is a free open-source high-level data manipulation library for Python that makes it very handy to work with structured or tabular data.
3. Matplotlib: a broad Python library for creating static, animated and interactive visualizations
4. Seaborn: a library for making attractive and informative statistical graphics in Python that builds on top of Matplotlib.
5. NumPy: A fundamental package in Python for scientific computing. It supports a large, multi-dimensional array and matrix plotting library as well as a collection of mathematical functions.
6. Scikit-learn (sklearn): a machine learning Python library that features pre-built tools for data processing, model selection, evaluation as well as implementing different learning algorithms.
7. SciPy: builds on NumPy and provides additional algorithms for optimization, linear algebra, integration, etc.

With these tools, we then conducted a rough array of analyses to assess the performance, consistency and alignment of LLM models with human evaluators on the website evaluation task. The results presentation is informed by the specific analyses.

The data underpinning the analysis reported in this paper are deposited at Figshare at <https://doi.org/10.6084/m9.figshare.28050476>.

## 4 Results and Analysis

Before discussing the performance evaluation result, it is important to note that we excluded the Mistral-7B-OpenOrca model from the analysis because it did not provide scores for several evaluated websites. This inconsistency in the model’s output will lead to bias and compromise the reliability of the comparative analysis. Therefore, we focused on the remaining models that consistently generated scores for all the websites under evaluation.

### 4.1 Model’s Internal Consistency Analysis

We assessed the consistency of each LLM model across three evaluation runs based on the total score and individual criteria scores. To illustrate the internal consistency of the LLM models, we calculated the standard deviation of the total scores across the three evaluation runs for each model and website. We visualized this analysis using a heatmap chart (Fig. 6).



Figure 6: Consistency Analysis of LLM Models in scoring four websites over three evaluation runs

The chart in Fig. 6 shows that most LLM models exhibited high consistency, with standard deviations close to zero for all websites. Based on this analysis, most models produced consistent total scores across multiple runs when evaluating a website. It means that most of the LLM models are reliable for website evaluation.

However, there were a few exceptions. The gpt-3.5-turbo-1106 model showed a relatively higher standard deviation of 4.64 for the "https://www.halalnavi.com" website, suggesting some inconsistency in its evaluations for that particular site. Additionally, the Mixtral-8x7B-Instruct-v0.1 model also had slightly higher standard deviations of 1.23 for

the "https://muslimguide.jnto.go.jp/eng" and 1.30 for the "https://www.halalmedia.jp" website. Another model, Gemini Pro, also showed slightly higher standard deviations of 1.04 for the "https://www.havehalalwilltravel.com" website.

To provide a broader perspective on evaluation consistency, we conducted a comparative variance analysis between LLM and human evaluations. Table 3 presents the variance statistics for each evaluation criterion across both groups. The analysis reveals several important patterns in evaluation reliability. First, LLM evaluations demonstrate systematically lower variance than human assessments across most criteria. For instance, in evaluating Credibility, LLMs show notably lower variance (5.481) compared to human evaluations (9.004), suggesting more consistent judgment in assessing source trustworthiness. Similar patterns are observed in Content Quality assessment, where LLM variance (2.469) is less than half of human variance (5.028), and in Relevance evaluation (LLM: 2.839 vs. human: 4.194).

Criteria	Human Evaluations		LLM Evaluations	
	Variance	Std Dev	Variance	Std Dev
Credibility	9.004	3.001	5.481	2.341
Relevance	4.194	2.048	2.839	1.685
Content Quality	5.028	2.242	2.469	1.571
Coverage	1.195	1.093	1.506	1.227
Comprehensiveness	0.879	0.938	0.620	0.788
Accessibility	3.577	1.891	1.042	1.021

Table 3: Comparison of Variance Statistics between Human and LLM Evaluations

Interestingly, both LLM and human evaluations demonstrate a logical hierarchical pattern in variance distribution. Technical criteria such as Coverage (LLM: 1.506, human: 1.195) and Comprehensiveness (LLM: 0.620, human: 0.879) show lower variance compared to more subjective criteria like Credibility and Content Quality. This consistency in patterns across both groups suggests that our evaluation framework successfully captures the inherent complexity of different criteria - more subjective criteria naturally show higher variance, while technical criteria demonstrate more stable assessments.

The lower variance in LLM evaluations, while maintaining these logical assessment patterns, provides strong evidence for the reliability of our automated approach. It suggests that LLMs can provide more consistent evaluations than human experts while preserving the ability to discriminate between different aspects of source quality. This finding is particularly important for ensuring reliable and scalable knowledge source evaluation.

In order to focus on other minor aspects of the total score consistency that can be improved, we took a closer look into the scoring patterns of each LLM model across the individual criteria for all of the websites using the LLMs. The standard deviation of the scores for various other criteria across all of the websites is illustrated in Fig. 7.

This analysis revealed that some models, such as GPT-3.5, Mixtral-8x7B, gpt-4, and Gemini-pro exhibited variations in their scores for specific criteria across different websites. Such variations leads models to be more responsive towards attributes and quality of each website individually in terms of its credibility, relevance, the quality of its content among many of other parameters evaluated. An evaluation tool should have

	Model	Credibility	Relevance	Content Quality	Coverage	Accessibility	Comprehensiveness
0	CodeLlama-34b-Instruct-hf	0.00	0.00	0.00	0.00	0.00	0.00
1	Mistral-7B-Instruct-v0.1	0.00	0.00	0.00	0.00	0.00	0.00
3	NeuralHermes-2.5-Mistral-7B	0.00	0.00	0.00	0.00	0.00	0.00
4	gemini-pro	0.00	0.00	0.00	0.00	0.26	0.00
6	gpt-4	0.00	0.00	0.00	0.06	0.06	0.00
7	zephyr-7b-beta	0.00	0.00	0.00	0.00	0.00	0.00
2	Mixtral-8x7B-Instruct-v0.1	0.22	0.17	0.24	0.00	0.00	0.00
5	gpt-3.5-turbo-1106-turbo-1106-turbo-1106	0.57	0.35	0.26	0.12	0.13	0.12

Figure 7: Consistency Analysis of LLM Models across criteria

the capacity to discern differences in the quality of the websites. This is possible by this tool, as it allows the models to be able to assess all the qualified sites' strengths and weaknesses.

## 4.2 Comparative Analysis

We conducted several comparative analysis methods, as follows:

### 4.2.1 Average Total Score Comparison

To determine the similarity or alignment between the LLM models and human evaluators, we compare their average total scores. Table 4 lists this comparison results. Based on these calculations, the Mistral-7B-Instruct-v0.1 and Mixtral-8x7B-Instruct-v0.1 models have the smallest differences from the human evaluators' average scores, at 2.29 and 2.71, respectively. Therefore, these two models are more closely aligned with human judgment.

Evaluator	Average Total Score	Absolute Difference
Human	79.28	0
CodeLlama-34b-Instruct-hf	83.89	4.61
Mistral-7B-Instruct-v0.1	76.99	2.29
Mixtral-8x7B-Instruct-v0.1	76.57	2.71
NeuralHermes-2.5-Mistral-7B	83.95	4.67
Gemini-pro	84.55	5.27
Gpt-3.5-turbo-1106	82.13	2.85
Gpt-4	83.77	4.49
Zephyr-7b-beta	96.19	16.91

Table 4: Average Total Score Comparison Between the LLMs and Human Evaluators

Meanwhile, the Gpt-3.5-turbo-1106 model, with a difference of 2.85, is the next most aligned with human evaluators, followed by Gpt-4, CodeLlama-34b-Instruct-hf,

and NeuralHermes-2.5-Mistral-7B. The Gemini-pro and Zephyr-7b-beta models have the most significant differences from the human evaluators' scores, at 5.27 and 16.91, respectively, indicating that their judgments are less aligned with human evaluators than the other LLM models.

#### 4.2.2 Website Rankings Comparison

In the second analysis, we compare the rankings assigned to the evaluated websites by the LLM models and human evaluators. Fig. 8 presents a heatmap chart that visualizes this comparison, with the numbers in each box indicating the rank. Lower ranks signify higher quality based on the evaluation criteria.

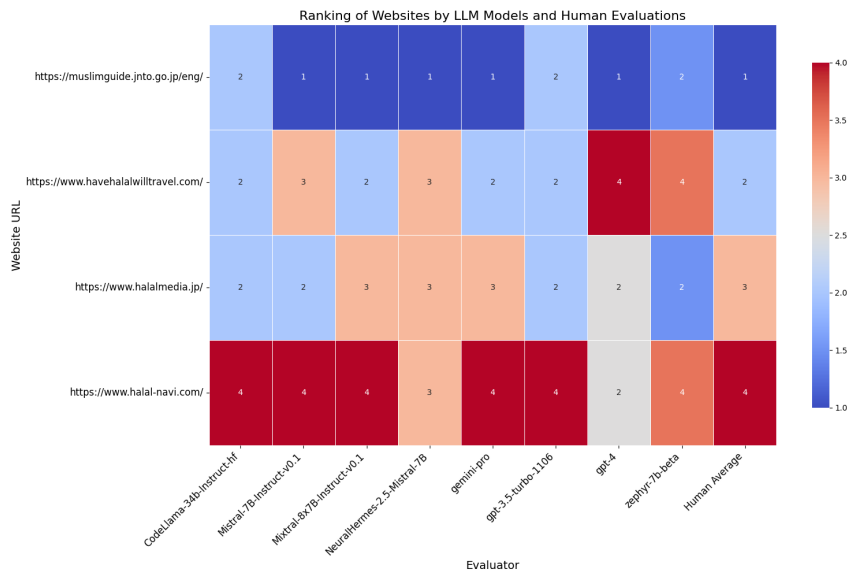


Figure 8: Website Rankings Comparison Between the LLM Models and Human Evaluators

According to the human evaluations, <https://muslimguide.jnto.go.jp/eng/> emerged as the top-ranked website, followed by <https://www.havehalalwilltravel.com/> in the second position. <https://www.halalmedia.jp/> and <https://www.halal-navi.com/> secured the 3rd and 4th ranks, respectively.

As we can see in Fig. 8, the Mistral-8x7B-Instruct-v0.1 and the Gemini-Pro models demonstrate the strongest alignment with the rankings provided by human evaluators. The CodeLlama-34b-Instruct-hf and Gpt-3.5-turbo-1106 models also exhibit a notable degree of alignment with human evaluators, particularly for two websites: <https://www.havehalalwilltravel.com/> and <https://www.halal-navi.com/>.

Furthermore, the Mistral-7B-Instruct-v0.1 and NeuralHermes-2.5-Mistral-7B models align with human evaluators for two websites each. The Mistral-7B-Instruct-v0.1 model matches the human rankings for <https://muslimguide.jnto.go.jp/eng/> and <https://www.halal-navi.com/>.

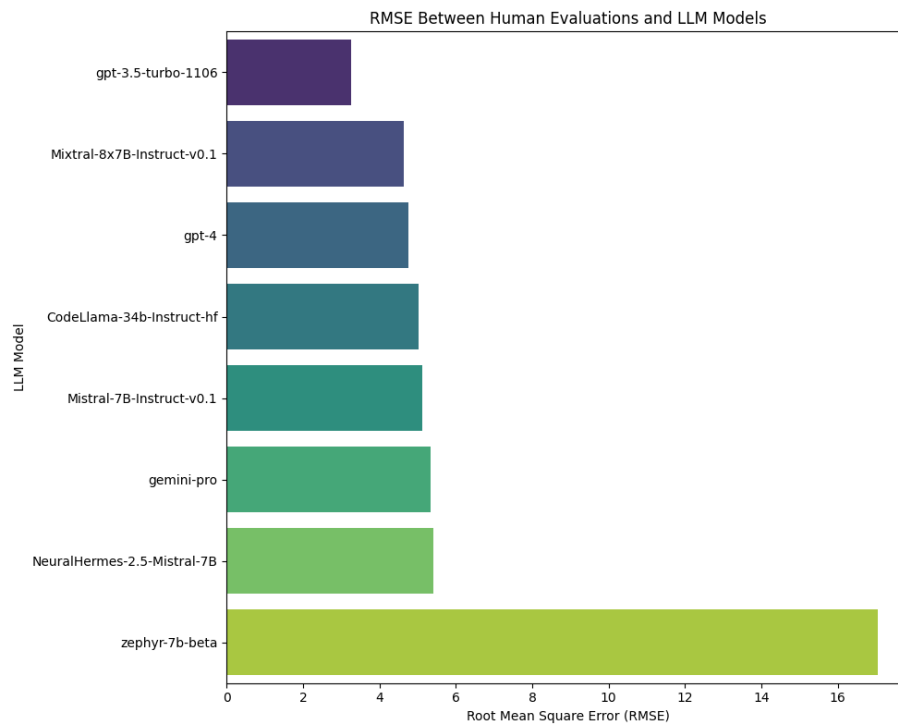


Figure 9: RMSE Analysis Comparison Between the LLM Models and Human Evaluators

halal-navi.com/, while the NeuralHermes-2.5-Mistral-7B model aligns with human judgments for <https://muslimguide.jnto.go.jp/eng/> and <https://www.halalmedia.jp/>.

Among the remaining models, GPT-4 demonstrates alignment with human evaluators only for the website <https://muslimguide.jnto.go.jp/eng/>. Its rankings for the other websites diverge from those assigned by the human experts. Notably, the Zephyr-7b-beta model exhibits the least alignment with human evaluators, with its rankings deviating significantly from the human-assigned ranks across all four websites.

#### 4.2.3 Root Mean Square Error Analysis

To further assess the agreement between the human evaluations and the LLM models, we calculated the Root Mean Square Error (RMSE) for each model. RMSE is a widely used metric that measures the average magnitude of the differences between predicted values (in this case, the LLM scores) and observed values (the human evaluation scores). Lower RMSE values indicate better agreement and closer alignment between the LLM models and human evaluators.

As shown in Fig. 9, the Gpt-3.5-turbo-1106 model has the lowest RMSE value (3.25), indicating the strongest agreement with the human evaluations. The Mixtral-8x7B-Instruct-v0.1 model had the second-lowest RMSE, followed by the gpt-4 model. These models also strongly agreed with the human evaluations, with RMSE values below 5.

The CodeLlama-34b-Instruct-hf, Mistral-7B-Instruct-v0.1, Gemini-pro, and NeuralHermes-2.5-Mistral-7B models have slightly higher RMSE values, around 5. Even though these values are higher than the top-performing models, they still indicate reasonable agreement with the human evaluations. Lastly, the zephyr-7b-beta model has the highest RMSE value, around 17, indicating the most significant deviations from the human evaluations among the tested models.

#### 4.2.4 Correlation Analysis

In addition to the RMSE metric, we employed the Spearman correlation method to examine the relationship between the human evaluations and the LLM models' scores across different evaluation criteria. The Spearman correlation is a non-parametric measure that assesses the strength and direction of the monotonic relationship between two variables. It is particularly useful when the data is ordinal or the relationship between variables is not necessarily linear [Scheff 2016].

Fig. 10 presents a heatmap of the Spearman correlation coefficients between the human evaluations and the LLM models for each evaluation criterion. The color intensity represents the strength of the correlation, with darker shades indicating stronger correlations.

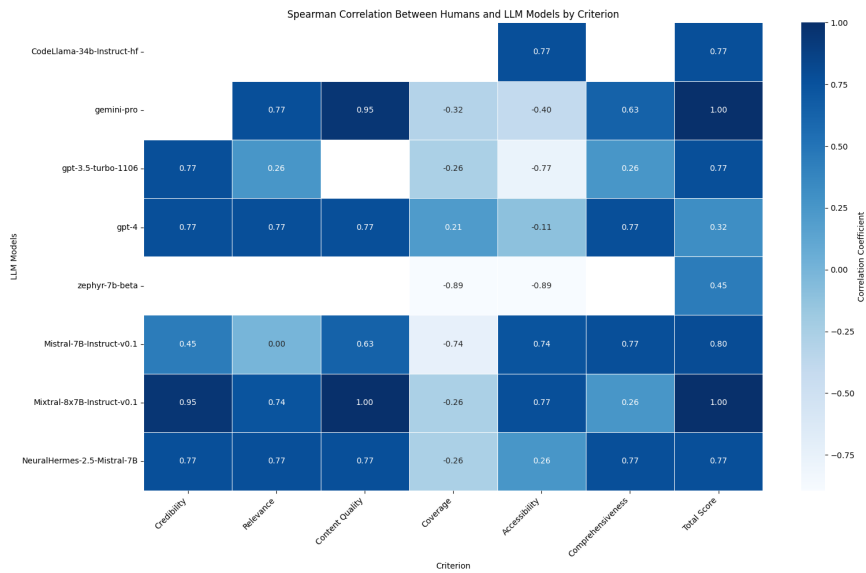


Figure 10: Spearman Correlation Matrix

Overall, the heatmap reveals a generally positive correlation between the human evaluations and the LLM models across most criteria. However, the strength of the correlations varies among the LLM models and criteria.

Models such as Mixtral-8x7B-Instruct-v0.1, gpt-4, and NeuralHermes-2.5-Mistral-7B show stronger correlations with human assessments than other models across multiple

criteria. However, the performance across different criteria is uneven for some models with others performing poorly or even negatively across some criteria such as coverage and accessibility.

Furthermore, Gemini-pro, Mistral-7B-Instruct-v0.1, and Gpt-3.5-turbo-1106 exhibited different degrees of correlation with human evaluation. These models have been categorized as exhibiting strong moderate and weak correlations with with certain criteria and in some cases missing data altogether. For example, Gemini-pro and Mistral-7B-Instruct-v0.1 proved to have a stronger positive correlation with content quality, while Gpt-3.5-turbo-1106 showed a greater central tendency toward human evaluations of credibility. However, those models also exhibit weak or negative correlations with areas like coverage or accessibility. Gpt-3.5-turbo-1106 has no data for content quality, and Gemini-pro has no trustworthiness coefficient.

Meanwhile, models like Zephyr-7b-beta and CodeLlama-34b-Instruct-hf, have many missing data and have strong negative correlations indicating a low capacity to replicate human evaluation trends accurately.

When examining the specific criteria, we observe that the content quality column has the darkest shades overall, indicating that most LLM models align well with human judgments when assessing the quality of website content. The credibility, relevance, and comprehensiveness columns also show relatively dark shades for several models, suggesting good agreement with human evaluations in these aspects.

On the contrary, the coverage column stands out as having the lightest shades and some strong negative correlations, particularly for the zephyr-7b-beta and Mistral-7B-Instruct-v0.1 models. This suggests that the LLM models generally struggle to accurately capture human judgments when evaluating the geographical coverage of websites. The accessibility column also shows a mix of positive and negative correlations, with models like CodeLlama-34b-Instruct-hf, Mistral-7B-Instruct-v0.1, and Mixtral-8x7B-Instruct-v0.1 demonstrating strong positive alignment. Meanwhile Gemini-pro, gpt-3.5-turbo-1106, gpt-4, and zephyr-7b-beta and exhibit negative correlations, indicating inconsistencies in assessing website accessibility compared to human evaluators.

#### 4.2.5 Statistical Differences Analysis

To statistically quantify the agreement in evaluations by LLM models and human assessments, we used the Mann-Whitney U test. This type of non-parametric statistical method is the best when there is a need to compare two independent samples especially for small sample size or data which is not normally distributed.

For each model and criterion combination, we calculated the p-values using the Mann—Whitney U test. When the p-value was less than 0.05 — below the standard threshold, that means the scores from LLM models differ the score of human test noticeably.

As shown in Fig.11, the heatmap chart demonstrating the p-values from Mann-Whitney U (MWU) test between LLM models' scores and human evaluators for each evaluation criterion. The darker shades in the heatmap represent higher p-values, indicating a stronger alignment between the LLM models and human evaluators. Lighter shades, on the other hand, correspond to lower p-values, suggesting more significant differences between the model scores and human judgments.

Overall, the heatmap reveals a generally positive picture, with most LLM models showing high p-values (darker shades) across the majority of the evaluation criteria. This indicates that the models' assessments are well-aligned with human judgments in most cases.

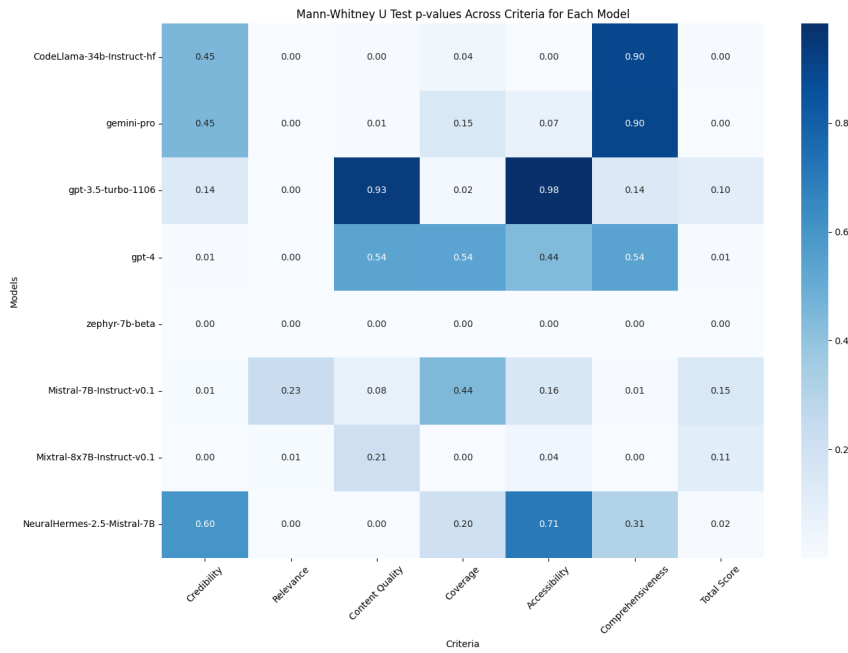


Figure 11: Mann-Whitney U Test p-Values Across Criteria for Each Model

Models such as gpt-4, gpt-3.5-turbo-1106, and NeuralHermes-2.5-Mistral-7B demonstrate particularly strong agreement with human evaluators, as evidenced by the consistently dark shades across several criteria. Other models, like Mistral-7B-Instruct-v0.1 and Gemini-pro, also show good overall alignment with human scores, with only a few lighter squares indicating minor deviations in specific criteria.

On the other hand, the CodeLlama-34b-Instruct-hf model has a mix of high and relatively lower p-values, suggesting that while they agree with human judgments in some aspects, there are areas where their assessments differ more notably. The zephyr-7b-beta and Mixtral-8x7B-Instruct-v0.1 models stand out as having the lowest p-values (lightest shades) across all criteria, indicating that their scoring patterns deviate the most from human evaluators compared to the other LLM models.

Looking at specific criteria, we see that credibility, accessibility, and comprehensiveness have the darker columns overall, suggesting that the LLM models are particularly good at aligning with human judgments in these aspects. On the other hand, the relevance criterion has the most lighter squares, indicating that most models have a lower agreement with humans when assessing the relevance of the websites. Only the Mistral-7B-Instruct-v0.1 model has no significance different from human evaluations in this criterion.

### 4.3 Overall Ranking of LLM Models

To gain insight from our comprehensive analysis, we derived an overall ranking to assess the relative effectiveness of the LLM models in the website evaluation task. We

employed a comprehensive scoring system, taking into account their performance across various evaluation analyses. For each analysis, we assigned scores based on the LLMs' relative performance, with lower scores indicating better results.

The scoring process was as follows:

1. For each evaluation analysis, we ranked the LLM models in order of their performance, with the best-performing model receiving a score of 1, the second-best receiving a score of 2, and so on.
2. In cases where multiple models exhibited similar performance or tied for a position, we assigned them the same score. For instance, in the comparison analysis based on the ranking result, the Gemini-pro and Mixtral-8x7B-Instruct-v0.1 were the top performers, so both received a score of 1.
3. We then summed the assigned scores for each model across all the evaluation analyses, providing a cumulative score that reflects their overall performance.
4. To obtain the final ranking, we divided the cumulative scores by the total number of evaluation analyses considered (in this case, 7), yielding an average score for each model.
5. We then ranked the LLM models based on their average scores, with lower average scores indicating better overall performance.

The scoring and ranking methodology has its advantage in making it possible to evaluate LLM models' capabilities in terms of the consistency, agreement with humans, and their performance across one or more statistics. By assigning scores based on relative performance and calculating an average score across multiple evaluation dimensions, the final ranking provides a balanced and holistic view of each model's capabilities in the website evaluation task.

The ranking of the LLM models is as follows:

1. Gpt-3.5-turbo-1106 and gpt-4 (Score: 2.71)
2. Mixtral-8x7B-Instruct-v0.1 (Score: 2.86)
3. Mistral-7B-Instruct-v0.1 (Score: 3.14)
4. NeuralHermes-2.5-Mistral-7B (Score: 3.86)
5. CodeLlama-34b-Instruct-hf (Score: 4.29)
6. Gemini-pro (Score: 4.43)
7. Zephyr-7b-beta (Score: 6.00)

## 5 Discussion

### 5.1 Key Findings and Implications

Our comparative analysis of LLM models and human evaluators uncovers multiple key findings with potentially high impact on knowledge graph creation. Models such as Gpt-3.5-turbo-1106, Gpt-4 and Mixtral-8x7B-Instruct-v0.1 are highly correlated with

human evaluations of the quality of online knowledge sources. These models can easily understand all the complexities involved in judging website credibility, relevance, and quality of content among other things.

In addressing potential hallucination concerns with LLM-based evaluation, our methodology incorporates several key safeguards. We designed our prompts to require simple numerical scores without explanatory text, reducing opportunities for unfounded assertions. Additionally, we employed a temperature setting of 0.4 across all models to balance output consistency with the ability to detect genuine content differences. Our system architecture further mitigates hallucination risks by having LLMs evaluate only actual scraped website content. The effectiveness of these strategies is evidenced by the low standard deviations in our consistency analysis (Figure 5), where most models demonstrated stable and reliable scoring behavior across multiple evaluation runs.

However, we caution that our analysis also suggests some caveats and nuances in the behavior of LLMs. All of them are very highly correlated with the level of expertise except for models such as Zephyr-7b-beta, Gemini-pro, and CodeLlama-34b-Instruct-hf, which in most cases have weaker correlations and much higher absolute deviation from human evaluation. Our results highlight the necessity for a truly informed choice and optimization of LLM models specific to evaluation tasks in order to tailor them perfectly to human judgments and performance.

Compared to previous research, such as the cost-based approach proposed by [Asgari et al. 2019], our LLM-based automated evaluation tool has several advantages. Our method leverages the high-level natural language processing features of large language models (LLMs) to process structured and unstructured data sources efficiently — a significant improvement over manual or semi-automated verification approaches. In addition, the adaptiveness of our approach to changes of knowledge and its capture of subtle semantic associations are what most traditional rule-based methods fail to handle.

## 5.2 Weighting System Analysis

Our evaluation framework utilizes a weight distribution that was developed through extensive human expert and LLM evaluation. The assigned weights reflect a dominance of essential quality measures (credibility: 30%, relevance: 24%, content quality: 21%) causes. The secondary quantities include coverage: 8%, comprehensiveness: 8% and accessibility: 9%. This hierarchy emphasizes an important principle in the construction of the knowledge graph – geographical breadth and technical ease of use are relevant but are rendered useless when the source is void of credibility and relevance. In this regard, statistical prove in the form of correlation analysis  $r > 0.9$  as well as hypothesis testing with  $p$  values  $> 0.8$  demonstrates a strong consensus between the assessments of the human experts and LLM models. This evidence once again strengthens the credibility of the weights [Hendrik et al. 2023].

So, while the most current distribution of the weights is acceptable and validated, there might be a specific implementation in the domain that would be specific. For instance, e-commerce knowledge graphs leverage more on higher accessibility weights to incorporate frequent price and inventory updates via APIs while highly delegating product information credibility requirements. Scientific research knowledge graphs may need a heavier reliance on content quality and credibility and a lesser emphasis on geographical reach but moderate accessibility on other attributes.

### 5.3 Implementation Consequences

It is important to consider the consequences that arise from implementing an LLM-based approach. First of all, the adoption of LLMs that require the use of commercial LLM APIs forces users to pose usage-based costs as well as dependency on certain providers. Given the relatively frequent transformations made to the model by LLM providers, it may be difficult to keep the evaluation standards at the same level over time, necessitating re-testing over time. In addition, processing a large amount of website content through LLM requires large computing resources, which could potentially affect the speed of evaluation on a large scale. All of these consequences should be taken into account whenever such systems are being deployed, especially in the case of large-scale knowledge graph construction projects.

### 5.4 Limitations

There are several limitations in our current study that need to be acknowledged. First, our comparative analysis was limited to a single case study of Halal tourism in Japan, which may affect generalizability to other domains or geographical contexts. This focused scope, while allowing for detailed analysis, leaves questions about performance across different knowledge domains and contexts.

Second, although we have developed an automated evaluation tool that significantly reduces the need for human coders, expert input remains necessary. This dependency manifests in two primary areas: developing domain-specific criteria and regular LLM model fine-tuning. Subject matter experts continue to play a crucial role in ensuring that evaluations remain relevant and appropriate for specific domains.

Third, our current system architecture, while effective for our case study, may face challenges when scaled to larger datasets or applied across multiple domains simultaneously. This limitation becomes particularly relevant when considering enterprise-scale knowledge graph construction projects.

### 5.5 Future Research Directions

Based on our findings and identified limitations, we suggest several key areas for further investigation that could significantly advance the field of automated knowledge source evaluation for knowledge graph construction.

#### 5.5.1 Model Enhancement and Integration

Future research should focus on advancing the core LLM evaluation capabilities by:

1. Development of ensemble approaches combining multiple LLM models, leveraging their complementary strengths in different evaluation criteria.
2. Creation of robust methods for integrating and validating new LLM models as they emerge, ensuring consistent evaluation standards.
3. Development of techniques for reducing model hallucination in source evaluation.

### **5.5.2 Weighting System Optimization**

Among major gaps that present opportunities for advancing the weighting framework include:

1. Development of context-aware weighting techniques that dynamically adjust based on domain characteristics and requirements.
2. Empirical validation that test the degree of effectiveness of different weights over various knowledge domains.
3. Study of automation techniques for weight optimization using domain-specific performance metrics.

### **5.5.3 System Architecture Enhancement**

To overcome scalability limitations, future work should explore:

1. Development of distributed architectures capable of handling large-scale source evaluation.
2. Employing intelligent caching and load balancing strategies to enhance the overall throughput.
3. Building fast and efficient parallel processing stacks of evaluating sources concurrently.
4. Investigation of fault-tolerant architectures for reliable operation.

### **5.5.4 Implementation and Resource Optimization**

Deployment of the system can be improved with respect to the following dimensions:

1. Formulation of policies to mitigate API dependencies and their cost.
2. Formulation of effective procedures for the LLM model update.
3. Optimization of computational resource usage.
4. Investigation of alternative approaches to reduce external service dependencies.

### **5.5.5 Cross-domain Validation and Automation**

Essential next steps for expanding the system's applicability involve:

1. In-depth research in various fields for the establishment of the universality of the approaches.
2. Creation of automated approaches that minimize expert requirements and yet do not compromise evaluation.
3. Development of autonomous systems for establishing and fine-tuning the criteria.
4. Development of methods facilitating the adjustment of domain specific optimization features to new situations automatically.
5. Development of techniques aimed at quick adjustment to new domains and data sources.

### 5.5.6 Comparative Evaluation Frameworks

The development of comprehensive comparative frameworks should involve:

1. Integration of the rule-based and classical NLP baseline methods as well as LLM evaluations in order to set performance indicators and appreciate the benefits of LLM integration.
2. Development of targeted assessment procedures able to evaluate the effectiveness of the traditional automated evaluation method against the LLM-based evaluation methods across structures, semi and unstructured types of web content.
3. The combination of rule-based approaches which will always highly be dependable with LLMs that are adaptable and intelligent will be sought.
4. Recruitment of benchmark datasets designed for the specific evaluation measures for different knowledge domains.
5. Comparative aspects of traditional approaches with LLM-based concepts in terms of computational costs and resources needed.

This research has demonstrated the potential of LLM-based automated evaluation tools in knowledge graph construction. The proposed future research directions intend to enhance this foundation, creating possibilities for establishing stronger, more scalable, and flexible systems for evaluating knowledge sources. As more capabilities are brought in the domain of knowledge graphs, it is expected that these research directions will be instrumental in achieving rich and trustworthy knowledge representation across domains.

## 6 Conclusion

This research proposes a new method for knowledge source assessment based on knowledge graph construction utilizing LLMs. This paper contributes and advances the field in three important areas. First, we illustrated the effectiveness of LLM-based tools in evaluations since their results strongly correlate with the judgments of human expert evaluators. This is particularly true when relying on models such as GPT-3.5-turbo, GPT-4, or Mixtral-8x7B-Instruct-v0.1. Secondly, we provided a detailed evaluation of the criteria with distinct weightings using an established criterion. Thirdly, we showed that various aspects of source quality including its credibility and relevancy, content quality, and accessibility can be evaluated by LLMs.

Our results have practical application in construction of knowledge graph. The automated source evaluation solution reduces the number of resources, time and effort required in evaluating sources, quality control is still enhanced. The ability of the system to handle both structured and unstructured data adds to its usefulness in real world problems. Additionally, our specific temperature level of 0.4 and nature of the prompt we used offers practical recommendations to users on how to use such systems.

In the bigger picture, this work expands the steps taken towards the construction of knowledge graphs with high reliability and scalability. In line with this trend, where more decision-making, innovations, and knowledge discovery are made with knowledge graphs, it is important that the necessity to incorporate automated processes for evaluating good quality sources becomes the first priority. Our work not only addresses this pressing requirement but also lays the groundwork for subsequent advances in automated

assessment of the quality of knowledge sources. Overall as the language models advance, there lies the opportunity for more advanced and better performing source evaluation systems which are also likely to further our capacities in knowledge representation in various fields to a greater level.

### Acknowledgements

We would like to express our gratitude to the anonymous reviewers for their insightful comments and constructive suggestions, which have significantly improved the quality of this manuscript. Their detailed feedback helped us strengthen our methodology and clarify various aspects of our research.

### References

- [Abu-Salih 2021] Abu-Salih, B.: "Domain-specific knowledge graphs: A survey"; *Journal of Network and Computer Applications* 185 (2021).
- [Asgari et al. 2019] Asgari-Bidhendi, M., Hadian, A., Minaei-Bidgoli, B.: "Farsbase: The Persian Knowledge Graph"; *Semantic Web* 10 (2019), 1169–1196.
- [Barrasa et al. 2021] Barrasa, J., Hodler, A., Webber, J.: "Knowledge Graphs"; O'Reilly (2021).
- [Brown et al. 2020] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: "Language Models are Few-Shot Learners"; *Proc. Neural Information Processing Systems, NIPS '20*, Curran Associates Inc., Red Hook, NY, USA (2020).
- [Chandak et al. 2023] Chandak, P., Huang, K., Zitnik, M.: "Building a Knowledge Graph to Enable Precision Medicine"; *Sci. Data* 10 (1) (2023), 67.
- [Chen et al. 2023] Chen, H., Jiao, F., Li, X., Qin, C., Ravaut, M., Zhao, R., Xiong, C., Joty, S.: "ChatGPT's One-Year Anniversary: Are Open-Source Large Language Models Catching Up?" (2023).
- [Chowdhery et al. 2024] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: "Palm: Scaling Language Modeling with Pathways"; *J. Mach. Learn. Res.* 24 (1) (2024).
- [Cascella et al. 2024] Cascella, M., Semeraro, F., Montomoli, J., Bellini, V., Piazza, O., Bignami, E.: "The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives"; *Journal of Medical Systems* 48 (2024).
- [Davis et al. 2024] Davis, J., Van Bulck, L., Durieux, B.N., Lindvall, C.: "The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research"; *JMIR Hum. Factors* 11 (2024).
- [Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding" (2019).

- [Faggioli et al. 2023] Faggioli, G., Dietz, L., Clarke, C.L.A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., Wachsmuth, H.: "Perspectives on Large Language Models for Relevance Judgment"; Proc. ACM SIGIR Int. Conf. Theory of Information Retrieval, ICTIR '23, ACM, New York (2023), 39–50.
- [Hendrik et al. 2023] Hendrik, H., Permanasari, A.E., Fauziati, S., Kusumawardani, S.S.: "Judging Knowledge by its Cover: Leveraging Large Language Models in Establishing Criteria for Knowledge Graph Sources Selection"; Proc. 8th Int. Conf. Information Technology and Digital Applications (ICITDA), 2023, 1–8.
- [Hogan et al. 2022] Hogan, A., Gutierrez, C., Cochez, M., de Melo, G., Kirrane, S., Polleres, A., Navigli, R., Ngomo, A.-C.N., Rashid, S.M., Schmelzeisen, L., Staab, S., Blomqvist, E., d'Amato, C., Gayo, J.E.L., Ncumair, S., Rula, A., Scudca, J., Zimmermann, A.: "Knowledge Graphs"; Springer International Publishing (2022).
- [Ji et al. 2022] Ji, S., Pan, S., Cambria, E., Martinen, P., Yu, P.S.: "A Survey on Knowledge Graphs: Representation, Acquisition and Applications"; IEEE Transactions on Neural Networks and Learning Systems 33 (2022), 494–514.
- [Jiang et al. 2023] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.-A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: "Mistral 7B" (2023).
- [Kasner and Dušek 2024] Kasner, Z., Dušek, O.: "Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-Text Generation" (2024).
- [Kuang et al. 2024] Kuang, S., Zhang, J., Mohajer, A.: "Reliable information delivery and dynamic link utilization in MANET cloud using deep reinforcement learning"; Transactions on Emerging Telecommunications Technologies 35(9) (2024).
- [Lenat and Marcus 2023] Lenat, D., Marcus, G.: "Getting from Generative AI to Trustworthy AI: What LLMs Might Learn from Cyc" (2023).
- [Langer et al. 2018] Langer, A., Siegert, V., Göpfert, C., Gaedke, M.: "SemQuire – Assessing the Data Quality of Linked Open Data Sources Based on DQV"; Proc. Current Trends in Web Engineering, Springer International Publishing, Cham (2018), 163–175.
- [Liu et al. 2019] Liu, Y., Zeng, Q., Ordieres Meré, J., Yang, H.: "Anticipating Stock Market of the Renowned Companies: A Knowledge Graph Approach"; Complexity 2019 (2019).
- [MacFarland and Yates 2016] MacFarland, T.W., Yates, J.M.: "Introduction to Nonparametric Statistics for the Biological Sciences Using R"; Springer International Publishing (2016), 103–132.
- [Mohamed et al. 2020] Mohamed, S.K., Nounu, A., Nováček, V.: "Biological Applications of Knowledge Graph Embedding Models"; Briefings in Bioinformatics 22, 2 (2020).
- [Naismith et al. 2023] Naismith, B., Mulcaire, P., Burstein, J.: "Automated Evaluation of Written Discourse Coherence Using GPT-4"; Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), ACL, Stroudsburg, PA, USA (2023).
- [OpenAI et al. 2024] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, L., Kamali, A., Kanitscheider, I., Keskar, N.S., Khan,

T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, L., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: "GPT-4 Technical Report" (2024).

[Pan et al. 2024] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: "Unifying Large Language Models and Knowledge Graphs: A Roadmap"; *IEEE Transactions on Knowledge and Data Engineering* 36 (7) (2024), 3580–3599.

[Raiaan et al. 2024] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., Azam, S.: "A review on large language models: Architectures, applications, taxonomies, open issues and challenges"; *IEEE Access*, 12 (2024), 26839-26874.

[Rane et al. 2024] Rane, N., Choudhary, S., Rane, J.: "Gemini or ChatGPT? Capability, Performance, and Selection of Cutting-Edge Generative AI in Business Management"; *SSRN Electron. J.* (2024).

[Regino et al. 2022] Regino, A., Caus, R., Hochgreb, V., Reis, J.: "Knowledge Graph-Based Product Recommendations on E-Commerce Platforms"; *Proc. 14th Int. Joint Conf. Knowledge Discovery, Knowledge Engineering and Knowledge Management, SCITEPRESS* (2022).

[Saeidnia 2023] Saeidnia, H.R.: "Welcome to the Gemini Era: Google DeepMind and the Information Industry"; *Libr. Hi Tech News* (Dec. 2023).

[Scheff 2016] Scheff, S.W.: "Fundamental Statistical Principles for the Neurobiologist: A Survival Guide"; *Elsevier* (2016), 157–182.

[Singh et al. 2023] Singh, C., Askari, A., Caruana, R., Gao, J.: "Augmenting Interpretable Models with Large Language Models During Training"; *Nature Communications* 14, 1 (2023).

[Smalheiser 2017] Smalheiser, N.R.: "Data Literacy: How to Make Your Experiments Robust and Reproducible"; *Elsevier* (2017), 157–167.

[Sottana et al. 2023] Sottana, A., Liang, B., Zou, K., Yuan, Z.: "Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence-to-Sequence Tasks" (2023).

[Thoppilan et al. 2022] Thoppilan, R., Freitas, D.D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H.S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M.R., Doshi, T., Santos, R.D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., Le, Q.: "LaMDA: Language Models for Dialog Applications" (2022).

- [Touvron et al. 2023] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: "Llama 2: Open Foundation and Fine-Tuned Chat Models" (2023).
- [Wang et al. 2021] Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., Wu, T., Chen, H.: "Knowledge Graph Quality Control: A Survey"; *Fundamental Research* 1 (2021), 607–626.
- [Wang et al. 2024] Wang, Q., Li, W., Mohajer, A.: "Load-aware continuous-time optimization for multi-agent systems: toward dynamic resource allocation and real-time adaptability"; *Computer Networks* 250 (2024).
- [Yang et al. 2024] Yang, T., Sun, J., Mohajer, A.: "Queue stability and dynamic throughput maximization in multi-agent heterogeneous wireless networks"; *Wireless Networks* 30 (2024), 3229–3255.
- [Yu et al. 2023] Yu, H., Yang, Z., Pelrine, K., Godbout, J.F., Rabbany, R.: "Open, Closed, or Small Language Models for Text Classification?" (2023).
- [Zhong et al. 2023] Zhong, L., Wu, J., Li, Q., Peng, H., Wu, X.: "A Comprehensive Survey on Automatic Knowledge Graph Construction"; *ACM Computing Surveys* 56 (4) (2023), 94:1–94:62.