

Leveraging Large Language Models for Enhanced Quality Assessment of Nutrition and Health Dashboards

Elena N Naumova, Bingjie Zhou, Kyle M Monahan, Peter Nadel, Anastasia Marshak, Kenneth Chui

Submitted to: Journal of Medical Internet Research
on: April 29, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 25

 Figures 26

 Figure 1..... 27

 Figure 2..... 28

 Figure 3..... 29

 Multimedia Appendixes 30

 Multimedia Appendix 1..... 31

 Multimedia Appendix 2..... 31

Leveraging Large Language Models for Enhanced Quality Assessment of Nutrition and Health Dashboards

Elena N Naumova¹; Bingjie Zhou¹ PhD; Kyle M Monahan² MS; Peter Nadel² BS; Anastasia Marshak¹ PhD; Kenneth Chui³ PhD

¹ Friedman School of Nutrition Science and Policy Tufts University Boston US

² Tufts Technology Services Tufts University Medford US

³ 150 Huntington Ave Department of Public Health and Community Medicine Tufts University School of Medicine Boston US

Corresponding Author:

Elena N Naumova

Friedman School of Nutrition Science and Policy
Tufts University
150 Harrison Ave
Boston
US

Abstract

Background: Data dashboards have become an essential tool in food and nutrition surveillance, enabling integration, visualization, and dissemination of multi-faceted data. While dashboards enhance data accessibility and decision-making, inconsistencies in data quality, standardization, and responsible conduct, along with misleading visualizations, limit their reliability and usefulness. Large Language Models (LLMs) could offer opportunities to assist in dashboard evaluation, yet their effectiveness compared to human expert assessments remains uncertain.

Objective: In this study, we evaluate the potential of LLMs in assessing nutrition and health dashboards by comparing ChatGPT-generated evaluations with expert reviews. We examine the alignment and discrepancies in scoring across dashboards and key dashboard quality indicators.

Methods: We developed a structured evaluation framework based on the 4E principles—Evidence, Efficiency, Emphasis, and Ethics—comprising 45 criteria. Seven publicly available nutrition and health dashboards were selected for evaluation. ChatGPT-4o was prompted to assess dashboards using extracted textual and visual content, generating scores and justifications for each criterion. Results were compared to previously published expert evaluations, analyzing ranking consistency and differences in dashboards and indicator-specific criteria.

Results: ChatGPT-4o successfully generated scores and justifications in accordance with the instructions provided in the prompt we designed. ChatGPT-4o rankings were well aligned with expert evaluations for dashboards included in our study (Spearman correlation = 0.79). When comparing average scores across all dashboards for specific evaluation criteria, granularity and completeness had high consistency, with both AI and human experts assigning relatively lower scores. However, ChatGPT assigned lower scores for standardization and higher scores for responsible conduct compared to human experts.

Conclusions: ChatGPT-4o could offer structured, scalable dashboard evaluations, showing reasonable alignment with expert assessments, especially in objective criteria like readability and accessibility. However, inconsistencies in assessing standardization, platform capability, and ethical considerations reveal limitations in contextual reasoning. While LLMs can enhance evaluation efficiency, expert oversight remains essential for accuracy and depth. Future research should explore diverse dashboards, compare multiple LLM models, and integrate multimodal capabilities to better assess interactivity and visualization integrity.

(JMIR Preprints 29/04/2025:76654)

DOI: <https://doi.org/10.2196/preprints.76654>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/>

Original Manuscript

Original Paper

Leveraging Large Language Models for Enhanced Quality Assessment of Nutrition and Health Dashboards

Bingjie Zhou ¹, PhD; Kyle M Monahan ², MS; Peter Nadel ², BS; Anastasia Marshak¹, PhD; Kenneth Chui ³, PhD; Elena N. Naumova ¹, PhD

1. Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA, USA
2. Tufts Technology Services, Tufts University, Medford, MA, USA
3. Department of Public Health and Community Medicine, Tufts University School of Medicine, Boston, MA, USA

Corresponding Author:

Elena N. Naumova, PhD
Friedman School of Nutrition Science and Policy
Tufts University
150 Harrison Ave
Boston, MA, 02111
United States
Phone: 6177850624
Email: elena.naumova@tufts.edu

Abstract

Background

Data dashboards have become an essential tool in food and nutrition surveillance, enabling integration, visualization, and dissemination of multi-faceted data. While dashboards enhance data accessibility and decision-making, inconsistencies in data quality, standardization, and responsible conduct, along with misleading visualizations, limit their reliability and usefulness. Large Language Models (LLMs) could offer opportunities to assist in dashboard evaluation, yet their effectiveness compared to human expert assessments remains uncertain.

Objective

In this study, we evaluate the potential of LLMs in assessing nutrition and health dashboards by comparing ChatGPT-generated evaluations with expert reviews. We examine the alignment and discrepancies in scoring across dashboards and key dashboard quality indicators.

Methods

We developed a structured evaluation framework based on the 4E principles—Evidence, Efficiency, Emphasis, and Ethics—comprising 45 criteria. Seven publicly available nutrition and health dashboards were selected for evaluation. ChatGPT-4o was prompted to assess dashboards using extracted textual and visual content, generating scores and justifications for each criterion. Results were compared to previously published expert evaluations, analyzing ranking consistency and differences in dashboards and indicator-specific criteria.

Results

ChatGPT-4o successfully generated scores and justifications in accordance with the instructions provided in the prompt we designed. ChatGPT-4o rankings were well aligned with expert evaluations for dashboards included in our study (Spearman correlation = 0.79). When comparing average scores across all dashboards for specific evaluation criteria, granularity and completeness had high consistency, with both AI and human experts assigning relatively lower scores. However, ChatGPT assigned lower scores for standardization and higher scores for responsible conduct compared to human experts.

Conclusions

ChatGPT-4o could offer structured, scalable dashboard evaluations, showing reasonable alignment with expert assessments, especially in objective criteria like readability and accessibility. However, inconsistencies in assessing standardization, platform capability, and ethical considerations reveal limitations in contextual reasoning. While LLMs can enhance evaluation efficiency, expert oversight remains essential for accuracy and depth. Future research should explore diverse dashboards, compare multiple LLM models, and integrate multimodal capabilities to better assess interactivity and visualization integrity.

Keywords: Nutrition surveillance, Large language models, Data dashboard, Dashboard evaluation

Introduction

Evolving Role of Nutrition Surveillance

Food and nutrition surveillance is a critical component of public health surveillance, tracking population nutrition status and identifying emerging nutrition issues over time [1]. This ongoing monitoring provides the evidence needed to evaluate nutrition programs, guide health policies, and allocate resources effectively to address public health priorities [2]. More importantly, modern nutrition surveillance integrates diverse datasets, linking nutritional status, anthropometric assessments, and nutritional biomarkers to chronic disease and mortality databases, as well as socioeconomic, agricultural, and environmental factors [3]. This interdisciplinary approach captures the complex interactions of food systems, economics, climate, and policy in shaping population nutrition, supporting data-driven and adaptive public health strategies [4].

Dashboards in Nutrition Surveillance: Opportunities and Challenges

Data dashboards - web-based applications that integrate comprehensive data visualizations, narratives, descriptive summaries, and structured data tables or databases - are essential tools for hosting, synchronizing, and disseminating multidimensional nutrition and health indicators from nutrition surveillance [5]. These dashboards enhance the ability to describe, monitor, and track key insights, strengthening the link between surveillance data and actionable public health responses and, in turn, supporting researchers, policymakers, students, and public health advocates in data-driven decision-making [6–8].

As dashboards become integral to nutrition surveillance, ensuring their accuracy and reliability is crucial. While they enhance data accessibility and decision-making, poorly designed dashboards can misrepresent information through inaccurate data, misleading visualizations, a lack of standardization, or insufficient context, potentially leading to flawed policies and interventions [5]. As such, the judgment of dashboards from experts in nutrition, public health, and UI/UX (UI: user interface; UX: user experience) design, along with diverse users, remains essential for interpreting findings, identifying inconsistencies, and contextualizing results within broader socioeconomic and policy environments [9].

LLM Applications and Challenges in Data Dashboards

Recent advancements in artificial intelligence (AI), particularly large language models (LLMs), offer new opportunities to enhance the lifecycle of data dashboards, improving accessibility for individuals without extensive computer programming experience. Research has demonstrated that LLMs can automate key aspects of data visualization, including recommendations, generation, and interpretation. Specifically, LLMs can analyze data types and user objectives to recommend suitable visualization formats, enable natural language querying, generate code to create visuals with error-detection feedback through an LLM agent, and provide detailed descriptions and interpretations of chart figures [10–13]. Furthermore, LLMs can be integrated with interactive data dashboards to

provide user-friendly, chat-like interfaces that allow users to retrieve visual summaries, interpret trends, and suggest statistical techniques simply by entering questions or observations [14]. This application of LLMs with multi-modal analytics represents a new frontier in data visualization characterized by more adaptive, intuitive, and human-centered interactions.

While current applications of LLMs reduce technological barriers to creating and interacting with dashboards, increasing efficiency and enabling users to engage more fully with the visuals and data presented, several challenges persist. Dashboards can still be challenging to use due to the complexity of their rich content and diverse data formats. Furthermore, issues such as a lack of data standardization, insufficient documentation of data sources, unclear representativeness in data collection, and misleading visualizations can undermine dashboard effectiveness and reliability [5]. Although AI applications enhance dashboard functionality, their lack of transparency and excessive reliance on automated analysis can obscure inaccuracies, misrepresent findings, and perpetuate biases [15,16]. These challenges, complicated further by inconsistencies in data sources, lack of standardization, and complex data structures, make it even more difficult to ensure the integrity, accuracy, and reliability of the information presented in dashboards.

Leveraging LLMs for Structured Dashboard Evaluation

A key advantage of LLMs is their ability to adapt domain-specific knowledge, follow prompt instructions, and execute evaluation tasks efficiently. This capability makes them useful for structured assessments, such as literature reviews, where they can follow PRISMA guidelines to systematically identify, synthesize, and analyze relevant studies [17]. Similarly, LLMs can be leveraged for dashboard evaluation by applying domain expertise and standardized frameworks, such as the 4E principles -- Evidence, Efficiency, Emphasis, and Ethics -- to assess complex data dashboards [18,19]. The 4E principles were originally designed for efficient scientific graphical displays. In brief, Evidence relies on the content, results, or concepts that correspond to the information the graph conveys, and the content, results, or concepts are transferrable to a verbal description. Efficiency states that a graph should help to explain the results or concepts by taking advantage of visual perception and utilize high standards of readability, interpretability, and clarity; Emphasis states that a graph should be user-friendly, force the intended audience to note the unexpected, motivate questions, and clarify statements; Ethics involves achieving the truth and promoting essential values such as accuracy, precision, reliability, integrity, and inclusiveness at neither excess nor deficiency, but at the mean of these two.

Study Objectives

This study aims to leverage the capabilities of LLMs to evaluate dashboards within the nutrition, food, and health-related domains. We propose to advance the development of more robust, intuitive, and adaptable dashboards in the health and nutrition sector by systematically examining how LLMs can ingest and apply expert knowledge and assist in identifying strengths, weaknesses, and potential areas of improvement in dashboards, and comparing the evaluation with the human expert review to assess the validity of this approach. The findings will contribute to broader guidelines for using AI-driven approaches in dashboard development and evaluation, enhancing both the scientific understanding of data visualization and its applications in public health decision-making [20].

Methods

Methods described in the following sections are summarized in the flowchart in Figure 1.

Criteria Development

We developed a knowledge graph-like framework that maps the logical connections between core principles and the major components of dashboards. This framework demonstrates the

interdependence of elements within the 4E principles:

- Data and documentation in the dashboard support the Evidence.
- Data serves as the foundation and building block for visualizations and, in turn, improves the Efficiency of delivering the data.
- Visualizations and their narratives in the dashboard highlight and offer the Emphasis on the key findings.
- The integration of proper documentation, accurate data, and visualization, along with supported materials such as links and videos, ensures adherence to the Ethics.

Figure 1. Overview of the dashboard evaluation process.

We retained the original structure of the established criteria published by Zhou et al., organizing them under the 4E principles (Evidence, Efficiency, Emphasis, and Ethics). Each criterion was structured with multiple indicators and sub-indicators, covering key indicators such as goals and scope, data quality, platform capabilities, conflict of interest, and other related aspects (**Figure 2**). Building upon this framework above, we revised and expanded the criteria by incorporating more detailed descriptions to optimize them for LLM ingestion and processing. We added more weight to the Ethics principle by relocating criteria previously categorized under contact and communication indicators and introducing a new visualization integrity indicator with additional criteria.

Figure 2. Knowledge graph framework mapping core 4E principles and dashboard components.

We further refined and consolidated criteria under granularity, interactivity, and goals and scope to enhance clarity and efficiency. As a result of these refinements, the total number of criteria was reduced from 48 to 45 (Table 1). Finally, to facilitate machine-readability and ensure consistency, the criteria were stored in a standardized format (JSON). This format enables structured data representation, allowing the ChatGPT API to efficiently parse, process, and apply the criteria across evaluations without manual reformatting.

Table 1. Principles, indicators, and metrics for dashboard evaluation.

Principles	Indicators	Sub-Indicators	Metrics
Evidence	Platform Goals and Scope		1. Clarity of purpose: The dashboard clearly states its primary goals and objectives.
			2. Measurement of the outcome: The dashboard defines how it measures success or progress towards its claimed goals and these metrics are relevant and easy to understand.
			3. Relevance of information: Data and information in the dashboard sufficiently support the claimed goals of the dashboard. For example, there are sufficient description and reasonable justification provided for each indicator/variable.
	Data Quality	Transparency	4. Data source documentation: All the data presented in the dashboard are accompanied by a clear description of data sources/links/references and definition of variables.
			5. Data scope: The dashboard provides information on data on population coverage/representativeness, timeframe, and sampling design.
			6. Data collection methodology: The dashboard clarifies data collection methods: surveys, cohorts, census, surveillance, or estimated from statistical models.
		Standardization	7. The dashboard enables comparison of data across multiple dimensions: temporal (time), spatial (geographical locations), and across different indicators/variables.
			8. The dashboard provides clear information on measurements and calibration methods, if any. The measurement unit is clearly provided and is uniform for data in the same category.

			9. The dashboard specifies the geospatial granularity at which data can be viewed and analyzed, including global, national, regional (state/province/region), local (village/city/town), and down to the postal code level.
			10. The dashboard specifies the temporal granularity at which data can be viewed and analyzed, including daily, weekly, monthly, quarterly, and yearly.
		Granularity	11. The dashboard allows for data to be viewed at multiple geospatial levels and offers the ability to drill down into finer
			12. The dashboard allows for data to be viewed at multiple temporal levels and offers the ability to drill down within the chosen temporal scope.
			14. Data related to nutrition in the dashboard can be broken down by specific nutrition aspects, such as nutrients, individual food items, categorized food groups, dietary patterns etc.
			15. Missing data are uniformly denoted across the dashboard using standard labels such as "Missing", "NA", "None" or "Not available".
		Completeness	15. Missing data are uniformly denoted across the dashboard using standard labels such as "Missing", "NA", "None" or "Not available".
			16. The dashboard provides an explicit definition and explanations for the occurrence of any missing data.
			17. The dashboard quantifies and presents the amount and percentage of missing data for each variable, including temporal and geospatial references.
		Efficiency	Platform Capability
19. The dashboard includes a tutorial to walk users through the entire platform (e.g., informative pop-ups, a tutorial video).			
20. The dashboard includes sufficient reference material, publications that related to the topics in the platform,			
21. The dashboard has the flexibility to incorporate more data.			
22. The dashboard provides the last updated date.			
Visualization Quality	Readability		23. The dashboard provides adequate and appropriate visualizations to demonstrate different features of the data like data distribution, geographical distributions, time trends, quality, quantity listing, and relationship of the data. (Is the visualization made easier to read than a table?)
			24. The dashboard provides a sufficient description for each visualization with adequate details on the context of the data.
			25. The visualizations have clear titles, axis labels, figure legends, logical data units and intervals, and geographical boundaries, to avoid clutter and confusion.
			26. Colors used should effectively represent different values or categories of data, with a logical progression or clear distinctions between different data points.

		Interactivity	27. The visualizations use effective and accessible color schemes (including those with color vision deficiencies.) and font sizes.
			28. The visualizations are available in high resolution.
			29. The dashboard includes interactive features such as tooltips, zooming, and scrolling to enhance user engagement and data exploration.
			30. The dashboard includes dropdown menus with multiple indicators for selecting data sets or categories.
			31. The dashboard includes dropdown menus that allow selection based on different levels of detail (e.g., regional vs. national data), gender, or sub-population with different social-economic status.
			32. The dashboard provides a visualization download option that supports downloading visuals at specified zoom levels, with applied filters or highlighted features.
Emphasis	Platform Accessibility		33. The dashboard is within the top results in search engines (e.g., Google, Bing) and follow SEO guidelines (keywords present, text highlightable, fast loading).
			34. The platform is user-friendly and allows users to easily navigate and explore data and visualizations according to ADA website compliance (e.g., text readable, alt-text on hovering over images, headers present).
	Dashboard Comprehension		35. The dashboard visualizations effectively highlight key textual information, markers, reference lines, common trends, and patterns, including numerical summaries such as means, medians, outliers, and specific events that emphasize critical points or areas.
			36. The dashboard has a good storytelling strategy with a clear focus on the main points, ensuring that contextual information is provided either directly within the visualization or through accompanying text, which is essential for understanding the full narrative behind the data.
			37. There are clear and logical flows and connections between visualizations, multi-panel plots, and embedded descriptions, effectively illustrating the relationships between different visuals or variable categories.
			38. The dashboard offers meaningful comparisons between groups, time periods, geographic locations using visualizations.
Ethics	Platform	Conflicts of Interest	39. Funding sources and roles of funders are clearly stated.
			40. Roles of all contributors to the dashboard are clearly described.
		Responsible Conduct	41. The dashboard explicitly highlights responsibilities to the public, funders, research subjects, research team colleagues, and other statisticians or statistics practitioners according to ASA Guidelines.
	Visualization Integrity		42. The numbers in tables and figures are corresponding to each other. If any calculations or selections happen, make sure there is an explanation for the data manipulation.

		43. The method of data representation (e.g., proportional symbols, choropleth mapping, bar, lines) should be consistent and appropriate for the type of data (categorical, ordered, continuous) being displayed.
	Contacts and Communication	44. The dashboard includes contact information for questions, suggestions and feedback.
		45. The dashboard requests information from users and provides a password-protected working environment.

Dashboard Selection

We utilized the 13 dashboards from Zhou et al. examined in 2022 [5] to compare the new evaluation framework with the completed expert evaluations. An updated search was conducted to assess the current status of these dashboards. Among them, we found that four dashboards significantly changed. Specifically, the Global Dietary Database completely removed the visualizations from the platform; The National Institute of Nutrition Atlas in India had a completely new design; Joint Child Malnutrition Estimates and State of the World's Children dashboards have been updated from the year 2019 to 2023 with new visuals and design. Additionally, two dashboards were excluded (Local Burden of Disease and Vitamin A Supplementation interactive dashboard) as they could not be accessed through automated testing due to structural differences from other dashboards. They were embedded within a larger platform, where interactivity changed without a corresponding change in the website address, preventing the automated extraction process from capturing screenshots across multiple pages with different website links. Ultimately, seven dashboards were identified and prepared for data extraction.

Data Extraction

We automated the collection and extraction of dashboard content, focusing on images, text, and structural elements, using the Python library Selenium. To identify relevant content, we first extracted links directly related to the main dashboard link, ensuring they shared the same prefix. This approach allowed us to capture webpages most closely associated with the main dashboard. After compiling a list of related links, we used Selenium to automatically capture a screenshot of each page. To further analyze the structure of the websites, we extracted specific HTML (HyperText Markup Language) tags, including paragraphs (<p>), headers (<h1> for main titles, <h2> for subsections, and <h3> for sub-subsections), text from <div> elements, anchor <a> tags along with their text and hyperlink references, attributes, and <path> tags associated with Leaflet interactive maps. The extracted data from each dashboard included the following for every sub-link: (1) the link itself, (2) a screenshot of the page, and (3) the corresponding HTML tags. This process ensured a comprehensive understanding of the dashboards' content and structure.

Prompt Design and Execution

We utilized the OpenAI API to develop a systematic workflow for evaluating dashboard content. This approach eliminates the need to manually interact with the ChatGPT interface or manage batch files, providing a more streamlined and traceable solution through Python code. The workflow of the prompt involves parsing the task into two roles: system and user. The system role is designed to provide the model with specific instructions and domain knowledge, ensuring consistent task execution. The user role feeds the extracted dashboard data into the model for analysis. As part of the system role, we provided a task description emphasizing that the AI is acting as an expert in the nutrition and health domain. The prompt included the 4E principles (Evidence, Efficiency, Emphasis, and Ethics), explanations of the knowledge graph-like evaluation framework, and a detailed

breakdown of 45 evaluation criteria. For the user role, the extracted dashboard data was prepared and fed into the model. Textual data was extracted and processed from JSON files using the strip method to preserve the structure while converting the JSON into a readable format for the model. Images were converted into Base64-encoded strings before being passed into the model. The prompts for both roles were sent to the OpenAI API using Python. We set a fixed seed value and temperature parameter as 0 to ensure reproducibility and deterministic response. We used the GPT-4o model as it accepts both text and image input and produces structured text outputs.

Model Output Evaluation

To evaluate the model's outputs, we designed the prompt to instruct ChatGPT to provide a score and justification for each of the 45 evaluation criteria grouped under the 4E principles (Evidence, Efficiency, Emphasis, and Ethics). The mean scores for each dashboard were calculated and ranked to compare the overall dashboard rankings between ChatGPT-generated assessments and expert evaluations. Additionally, a detailed comparison of scores was conducted across all aligned evaluation criteria (sub-indicators) for each dashboard to identify patterns of agreement and divergence within the evaluation framework. To quantify the relationship between ChatGPT-generated and expert scores, the Spearman correlation was computed for each sub-indicator, evaluating the rank-based association between the two scoring methods. The percentage difference between expert and AI scores was calculated to measure the extent of discrepancies. To further assess agreement and identify potential biases, a Bland-Altman analysis was performed, examining the distribution of differences between AI and expert scores relative to their mean values. This analysis allowed for the detection of systematic over- or underestimation by ChatGPT and the presence of proportional bias, where discrepancies varied depending on the magnitude of the scores.

In addition to numerical scores, the model-generated justifications for each criterion were collected and subjected to qualitative analysis. This manual review focused on identifying key themes, common strengths, and limitations across the evaluated dashboards. The justifications were summarized to provide a detailed understanding of the model's reasoning and highlight areas where the LLM captured or missed critical aspects compared to expert evaluations.

The complete dataset and codes are available on GitHub: https://github.com/Bingjiezhou/dashboard_llm/.

Results

Criteria Development

A total of 45 criteria were developed and refined, with a primary focus on the principle of Evidence, especially data quality (14/45), and on the principle of Efficiency, particularly visualization quality (10/45). Other criteria covered aspects such as platform goals and scope, capability, accessibility, and so on. The full set of structured criteria is detailed in **Table 1**.

Dashboard Selection

The selected dashboards represent a diverse range of global nutrition and food security monitoring tools, each serving distinct yet complementary roles in nutrition surveillance and policy development. These platforms are created by government agencies, international organizations, and research institutions, providing insights into food systems, malnutrition, dietary intake patterns, and national and global nutrition policy progress. Some dashboards focus on tracking policy commitments and fortification programs, while others integrate multi-sectoral data on food security and individual consumption patterns. Additionally, real-time monitoring dashboards offer dynamic updates on hunger and malnutrition hotspots, supporting rapid response efforts. Evaluating these

dashboards allows for a comprehensive assessment of data quality, visualization effectiveness, and usability across different types of nutrition-focused dashboards. Detailed descriptions of selected dashboards are provided in Multimedia Appendix 1.

Data Extraction

The number of first-layer links extracted from each dashboard varied depending on the website's structure, ranging from a minimum of five to a maximum of 31 (mean: 15). While some dashboards organized their content across multiple sub-pages, others used tabs or embedded JavaScript to display interactive content within the same webpage link. Each extracted link was accompanied by a full-page screenshot and a text file containing the extracted key component from the HTML.

Prompt Design and Execution

To ensure a structured and replicable evaluation process, we implemented a systematic prompting strategy using the OpenAI API and the GPT-4o-2024-11-20 model. The API prompt defined the role of the model as an expert evaluator and established the evaluation framework. The user message included explicit instructions, a scoring system (1-10 scale), justification requirements, and evaluation materials. A summary of the API configuration is provided in Table 2.

Table 2. Configuration of the OpenAI API prompt for automated dashboard evaluation.

Prompt Component	Content
System Message	You are an expert evaluator of nutrition and health dashboards. Based on the definition, knowledge graph, and detailed criteria, evaluate the dashboard using the provided information.
Evaluation Framework	The evaluation follows the 4-E principle: detailed definition of the principles (see Table 1).
Knowledge Graph	The 4-E principle also supports a knowledge graph: - The data and documentation in the dashboard support the Evidence principle. - Data produces the visualization. - The visualization in the dashboard improves Efficiency principle. - The visualization and its narratives in the dashboard highlight Emphasis. - Proper documentation, accurate data, and visualization (together with supported links, videos, etc.) ensure Ethics principle.
Evaluation Criteria	The evaluation is based on the following criteria under the 4-E principle: a list of 45 criteria in detail
User Message - Instructions	Evaluate the dashboard by applying the above criteria. For each criterion, provide: - A score from 1 to 10 (Scores should be objective and justified based on the content). - A structured justification with specific references where applicable. Maintain a professional and analytical tone—avoid overly positive or generic statements. Clearly highlight both strengths and weaknesses.
Scoring Scale	1-4: Needs significant improvement (poor usability, unclear insights, or major issues). 5-6: Acceptable but with noticeable limitations (functional but not optimal). 7-8: Good to very good (effective but could be further refined). 9-10: Excellent to industry-leading (best practices in dashboard quality).
Assistant Response	Each score must be accompanied by a clear and structured explanation.

Guidelines	- Justifications should be evidence-based, detailed, and directly tied to the evaluation criteria. - The response should highlight areas for improvement while acknowledging strengths. - The assessment should remain neutral, analytical, and focused on objective evaluation.
Message Structure	- System Message: Defines the evaluation framework, scoring methodology, and criteria. - User Message: Directs the AI to provide structured, evidence-based responses.

Model Output Evaluation

Scoring Results

ChatGPT successfully produced a score and a justification. The detailed evaluation results are provided in Multimedia Appendix 2. We found that even if we set up a fixed seed value and 0 for the temperature parameter, the score and justification still varied a bit every time. So, we repeated the evaluation three times for each dashboard and calculated the mean for each criterion in each dashboard. In order to compare to the previously published evaluation results by experts, we ranked the dashboard based on the average score across all 45 criteria. Table 3 shows the comparison between the experts in 2022 and the ChatGPT evaluation in our study. The ranking is not the same, but similar (Spearman correlation = 0.79); the higher-end and lower-end dashboards are consistent, with each dashboard fluctuating plus or minus two places at most. This fluctuation could be due to the update of the dashboard between 2022 and 2025, the data extraction variation, and the evaluation quality by ChatGPT.

Table 3. Comparison of dashboard rankings: ChatGPT-4o vs. human evaluations.

Dashboard	Ranking	
	Experts	ChatGPT 4-o
Foods System Dashboard	1	2
Global Fortification Data Exchange	2	1
Global Individual Food consumption data Tool	3	4
Hunger Map	4	3
The State of Acute Malnutrition	5	7
Nutrition Landscape Information System	6	6
Global Nutrition Targets Tracking Tool	7	5

We further compared detailed scores given by ChatGPT and human experts for detailed indicator categories across all dashboards (Table 4). As in the previous study, we used a 5-point Likert scale of 1-low to 5-high; therefore, the expert scores were calibrated by multiplying them by 2 to ensure comparability with the ChatGPT-generated scores. The percentage difference was calculated by taking the difference between the ChatGPT-generated score and the expert score and dividing it by the expert score. Spearman correlation values indicated that ChatGPT scores align well with expert evaluations in structured and objective sub-indicators like Readability (0.77) and Platform Accessibility (0.88). Moderate to strong agreement was found for some sub-indicators, such as Platform Goals and Scope (0.67), Completeness (0.60), and Responsible Conduct (0.63). However, weaker correlations are observed in subjective areas like Platform Capability (0.25) and Contacts and Communication (0.17). ChatGPT tended to assign higher scores than human experts, particularly in

sub-indicators like Granularity (up to 80.3% difference), Completeness (up to 92.9%), and Platform Capability (up to 115.7%).

The Bland-Altman plot (Figure 3) illustrates the differences between ChatGPT and expert scores relative to their average values. The mean difference was slightly positive, indicating a general tendency for ChatGPT scores to be higher than expert scores in certain cases. Most data points fell within the range of the mean difference ± 1.96 times the standard deviation of differences. The plot also revealed a downward trend between the average and difference in scores, indicating the magnitude of the difference between ChatGPT and expert evaluations tended to decrease as the overall dashboard score increased. This suggests that ChatGPT-generated scores were more aligned with expert assessments for dashboards receiving higher overall ratings, whereas it tended to estimate lower-rated dashboards compared to human experts.

Table 4. Comparison of evaluation criteria score: ChatGPT-4o vs. human evaluation.

Principle	Indicator	Sub-indicator	Dashboard	Experts	ChatGPT -4o	Difference (%)		
Evidence	Platform Goals and Scope (ρ=0.67)		Foods System	9.58	8.67	-9.6		
			Global Fortification Data	9.58	8.67	-9.6		
			Global Individual Food	9.08	8.67	-4.6		
			Hunger Map	8.42	8.67	3.0		
			Acute Malnutrition	9.00	8.44	-6.2		
			Landscape Information	9.33	8.67	-7.1		
			Global Target Tracking Tool	7.58	8.33	9.9		
	Data Quality		Transparency (ρ=0.54)		Foods System	8.56	9.00	5.2
					Global Fortification Data	7.44	9.11	22.4
					Global Individual Food	8.00	7.44	-6.9
					Hunger Map	8.78	9.22	5.1
					Acute Malnutrition	6.11	7.67	25.5
					Landscape Information	7.11	8.33	17.2
					Global Target Tracking Tool	6.11	8.33	36.4
			Standardization (ρ=0.49)		Foods System	9.08	8.75	-3.7
					Global Fortification Data	8.17	9.33	14.3
					Global Individual Food	8.08	8.08	0.0
					Hunger Map	7.33	8.75	19.3
					Acute Malnutrition	7.75	7.92	2.2
					Landscape Information	7.67	8.08	5.4
					Global Target Tracking Tool	6.83	8.00	17.1
			Granularity (ρ=0.54)		Foods System	6.13	8.00	30.4
					Global Fortification Data	4.93	8.17	65.5
					Global Individual Food	7.33	7.75	5.7
					Hunger Map	6.00	7.83	30.6
					Acute Malnutrition	4.53	6.92	52.6
					Landscape Information	4.33	7.25	67.3
					Global Target Tracking Tool	4.07	7.33	80.3
			Completeness (ρ=0.60)		Foods System	4.83	8.00	65.5
					Global Fortification Data	4.67	9.00	92.9
					Global Individual Food	3.75	6.78	80.7
					Hunger Map	5.92	8.00	35.2
					Acute Malnutrition	4.25	6.78	59.5
					Landscape Information	5.67	7.22	27.5
					Global Target Tracking Tool	4.58	7.56	64.8
Efficiency	Platform Capability		Foods System	7.89	9.27	17.5		

	(ρ=0.25)		Global Fortification Data	9.22	9.60	4.1
			Global Individual Food	7.78	8.60	10.6
			Hunger Map	4.11	8.87	115.7
			Acute Malnutrition	5.67	7.93	40.0
			Landscape Information	8.44	7.87	-6.8
			Global Target Tracking Tool	5.56	7.93	42.8
	Visualization Quality	Readability (ρ=0.77)	Foods System	9.00	8.89	-1.2
			Global Fortification Data	7.93	9.17	15.5
			Global Individual Food	7.27	8.44	16.2
			Hunger Map	7.67	8.94	16.7
			Acute Malnutrition	7.20	8.33	15.7
			Landscape Information	7.20	8.17	13.4
		Interactivity (ρ=0.54)	Global Target Tracking Tool	5.73	8.39	46.3
			Foods System	9.00	9.00	0.0
			Global Fortification Data	8.39	9.42	12.3
			Global Individual Food	6.56	8.42	28.4
			Hunger Map	6.89	8.17	18.5
			Acute Malnutrition	8.61	7.92	-8.1
			Landscape Information	5.78	7.42	28.4
			Global Target Tracking Tool	6.33	7.92	25.0
Emphasis	Platform Accessibility (ρ=0.88)		Foods System	8.92	9.00	0.9
			Global Fortification Data	8.42	9.00	6.9
			Global Individual Food	7.17	8.17	14.0
			Hunger Map	8.00	8.67	8.3
			Acute Malnutrition	7.92	8.17	3.2
			Landscape Information	6.92	7.83	13.3
	Dashboard Comprehension (ρ=0.69)	Global Target Tracking Tool	6.83	8.17	19.5	
		Foods System	7.67	8.58	12.0	
		Fortification	8.20	8.92	8.7	
		Global Individual Food	6.27	8.00	27.7	
		Hunger Map	7.20	8.58	19.2	
		Acute Malnutrition	6.73	7.92	17.6	
		Landscape Information	6.40	7.83	22.4	
		Global Target Tracking Tool	5.47	8.00	46.3	
Ethics	Conflicts of Interest (ρ=0.41)		Foods System	7.00	10.00	42.9
			Global Fortification Data	7.50	9.67	28.9
			Global Individual Food	6.67	8.50	27.5
			Hunger Map	5.17	9.50	83.9
			Acute Malnutrition	6.50	8.50	30.8
			Landscape Information	5.50	8.33	51.5
	Responsible Conduct (ρ=0.63)	Global Target Tracking Tool	5.00	9.00	80.0	
		Foods System	9.33	9.33	0.0	
		Global Fortification Data	9.17	10.00	9.1	
		Global Individual Food	6.17	9.00	45.9	
		Hunger Map	6.50	9.67	48.7	
		Acute Malnutrition	7.83	9.00	14.9	
		Landscape Information	6.17	8.67	40.5	
		Global Target Tracking Tool	7.50	9.00	20.0	
	Contacts and Communication (ρ=0.17)	Foods System	8.33	8.67	4.0	
		Global Fortification Data	5.67	8.67	52.9	
		Global Individual Food	5.67	8.17	44.1	
		Hunger Map	7.33	8.50	15.9	
		Acute Malnutrition	6.67	7.67	15.0	
		Landscape Information	7.00	7.67	9.5	
Global Target Tracking Tool	6.67	8.33	25.0			

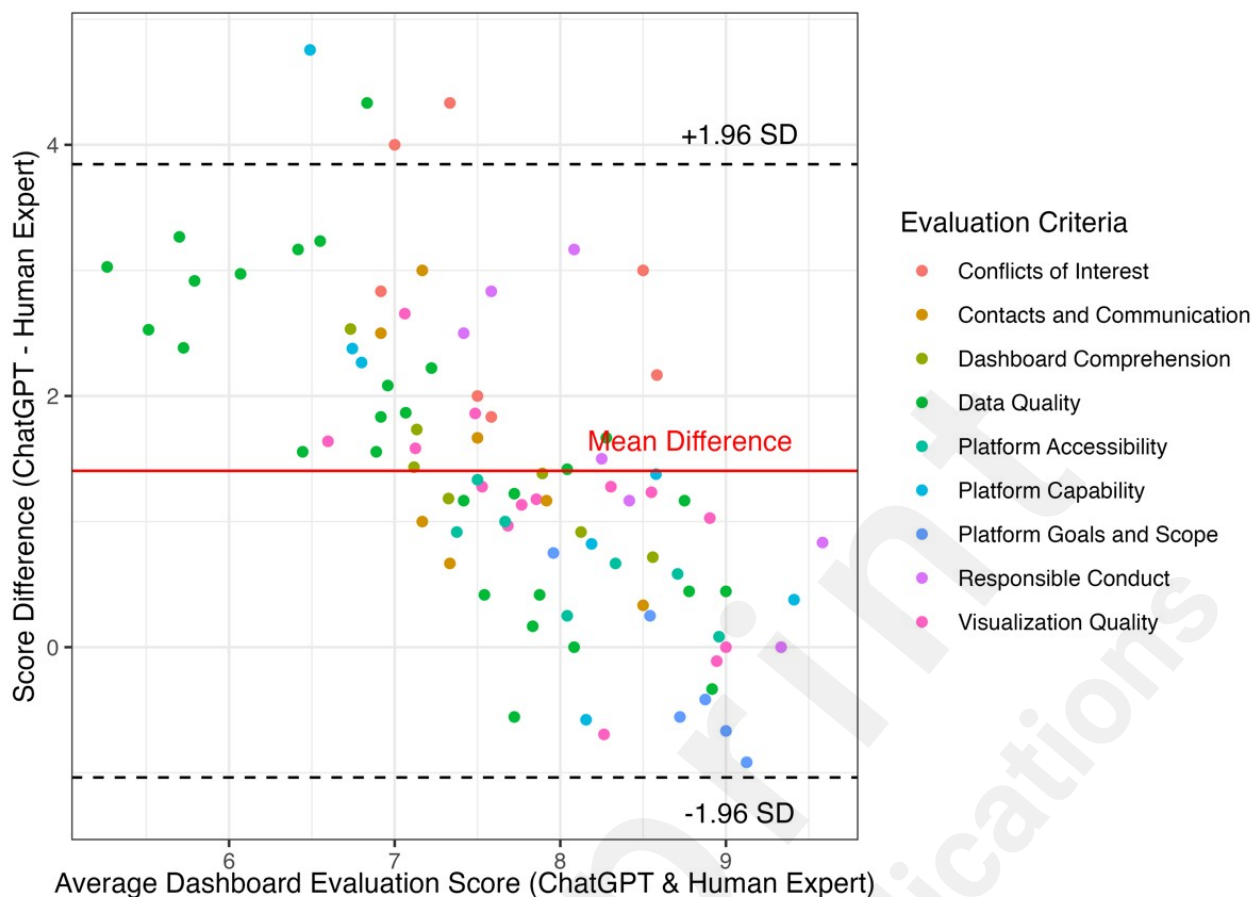


Figure 3. Agreement between ChatGPT-4o and experts in dashboard evaluation.

Justification of Evaluation Results

ChatGPT's justifications provided valuable insights into dashboard evaluations, referencing specific dashboard elements to support its assessments. For instance, for the "goals and scope" criterion, the model noted: *"The dashboard clearly states its purpose: to provide a comprehensive view of food systems by integrating data from multiple sources. This is evident in the 'About the Food Systems Dashboard' section."* Similarly, it provided relevant details for data exploration and accessibility, stating: *"Users can explore and download data in multiple formats, including Excel and Word."* However, in some cases, justifications lacked specificity. For example, for the readability criterion, the justification was overly general: *"Titles, axis labels, and legends are clear and well-structured."*

Discussion

This study presents a novel approach to evaluating nutrition and health dashboards by integrating LLMs into a structured assessment framework. By automating the evaluation process based on predefined criteria, extracting dashboard content programmatically, and applying ChatGPT-4o through the OpenAI API, we systematically assessed dashboard quality across multiple dimensions while reducing the need for manual review. Our findings demonstrate that ChatGPT-4o can generate structured evaluations that align with expert assessments across key indicators, reinforcing its potential as a scalable and efficient addition for assessing nutrition and health data dashboard platforms. However, our analysis also highlights areas of alignment and discrepancy between LLM-generated and expert evaluations. The Spearman correlation (0.79) for the dashboard ranking between ChatGPT and expert suggests a general agreement in identifying higher- and lower-quality dashboards, with fluctuations likely attributed to updates in dashboard content, variations in data extraction, and the inherent characteristics of AI-generated assessments. Additionally, when

examining specific evaluation criteria, we observed high agreement in structured and objective indicators, such as Readability (0.77) and Platform Accessibility (0.88), but weaker correlations in subjective and interpretive indicators, such as Platform Capability (0.25) and Contacts and Communication (0.17). These discrepancies suggest that ChatGPT struggles with more nuanced assessments that require contextual understanding, human judgment, or the ability to infer meaning beyond explicit textual cues. The percentage differences showed that ChatGPT-4o consistently assigned higher scores than human experts, meaning that ChatGPT tended to overestimate dashboard quality, especially in categories such as Granularity (up to 80.3), Completeness (up to 92.9), and Platform Capability (up to 115.7). The Bland-Altman analysis further contextualized these differences with a downward trend, indicating that ChatGPT's assessments were more aligned with expert reviews for higher-rated dashboards but overestimated lower-rated dashboards, suggesting that it may struggle to critically assess dashboards with incomplete documentation, weaker data structures, or poor standardization. This trend raises important considerations regarding AI-assisted evaluation reliability, particularly when applied to dashboards that require in-depth expert judgment.

Strengths and Limitations

The findings presented in this study highlight both the strengths and limitations of using ChatGPT-4o for dashboard evaluation. First, this study employs a well-defined evaluation framework based on the 4E principles (Evidence, Efficiency, Emphasis, and Ethics) [19], ensuring a structured assessment of dashboard quality informed by expert domain knowledge. Second, this study establishes an automated and structured approach that follows a sequential process, starting with data extraction and subsequently applying LLMs for evaluation. This approach enhances scalability and reproducibility while reducing the time and labor required for manual assessments, ensuring consistency across different dashboards. Additionally, comparing AI-generated evaluations with expert rankings offer valuable insights into the alignment and discrepancies between human and AI assessments in assessing text, data visualizations, and website structures. Similar studies in other domains have also compared ChatGPT-generated results with expert evaluations, further demonstrating AI's capabilities and limitations in complex decision-making tasks [21,22]. This analysis underscores the importance of critically interpreting AI-generated information. Finally, this study highlights key challenges and opportunities in AI-driven dashboard assessment, paving the way for future improvements in multimodal AI models for processing and interpretation of advanced data visualization and interactive dashboards [23].

A key limitation of this study is that the justifications provided by LLMs were often overly general and lacked transparency. Given the black-box nature of AI models, it remains difficult to trace how these textual responses were generated or to verify the evidence used in their reasoning [24]. This underscores the necessity of keeping human reviewers in the loop to validate AI-generated assessments, ensuring accuracy and contextual relevance. While LLMs can efficiently analyze structured criteria, human expertise remains essential in interpreting complex findings and identifying gaps in LLMs-generated justifications [25]. Another limitation of LLMs in dashboard evaluation is their inability to process interactive components effectively. Current AI models, including ChatGPT and other LLMs, struggle to navigate and interpret dynamic web elements. While they excel at extracting textual content, they cannot fully engage with interactive dashboards or understand how users interact with visualizations. Despite recent advancements in multimodal AI models that enable image-to-text processing, LLMs still have limited capacity to interpret visual data elements, such as trends, legends, color, and complex graphical representations. Human perception also faces challenges in deciphering intricate visualizations, but humans retain a critical advantage in identifying meaningful patterns, contextual cues, and inconsistencies that AI may overlook. As AI agent technologies advance, there is potential for LLMs to evolve into more sophisticated tools that can navigate web-based dashboards, recognize structural hierarchies, and analyze interactivity beyond static screenshots[26]. Future iterations of AI systems could incorporate adaptive browsing

capabilities, allowing them to dynamically interact with different types of dashboards in a more context-aware and structured manner.

Future Directions and Improvements

To further refine AI-assisted dashboard evaluation, future research should explore testing multiple LLMs, such as ClaudeAI, Copilot, and open-source models with enhanced image-to-text capabilities. Evaluating multiple LLMs will help assess their comparative strengths and weaknesses when analyzing dashboards. Comparing expert rankings with outputs from different AI models will provide a broader understanding of LLM reliability and bias in evaluation outcomes, offering insights into the extent to which AI-generated assessments align with human expertise. Expanding the number of dashboards assessed across diverse domains and design complexities will enhance the generalizability of findings, allowing for a more comprehensive evaluation of LLMs' performance. Additionally, incorporating control questions and expert-validated examples as benchmarks will strengthen model validation, ensuring the reproducibility and reliability of the evaluation. Beyond evaluation, future research could investigate whether ChatGPT and other LLMs can generate user-friendly tutorials for dashboard navigation and usage when provided with a dashboard. Specifically, testing whether AI can create example use cases, step-by-step guidance, and contextual explanations tailored to different user needs would offer valuable insights into its potential for enhancing user interaction, accessibility, and knowledge transfer. This type of analysis and evaluation could serve as a valuable tool for experiential learning to emphasize both the advantages and limitations of AI applications in reasoning tasks [20,27–30].

Conclusions

This study demonstrates the feasibility of LLM-assisted evaluation for nutrition and health dashboards, showing strong alignment between ChatGPT-4o and expert assessments in structured, rule-based criteria. However, observed biases, overestimations in specific categories, and discrepancies in subjective assessments highlight the importance of integrating AI with human expertise. While LLMs offer a scalable, efficient tool for evaluating dashboards, expert oversight remains crucial to ensure accuracy, reliability, and interpretative depth. Future AI-driven evaluation frameworks should focus on enhancing multimodal capabilities, refining scoring models, and improving AI explainability, paving the way for more comprehensive and adaptive dashboard assessment methodologies.

Authors' Contributions

BZ led the study design, methodology, and manuscript writing. KMM and PN contributed to the study design and methodology. KC and AM assisted with manuscript revisions and provided critical feedback. ENN supervised the study and guided the research framework.

Acknowledgement

This study was supported by the National Science Foundation Innovations in Graduate Education Program grant (Award #1855886, PI Naumova) titled SOLution-oriented, STudent-Initiated, Computationally Enhanced (SOLSTICE) Training and Société des Produits Nestlé SA. The funding source had no involvement in the study design; the collection, analysis, and interpretation of data; the writing of the report; and the decision to submit the manuscript for publication.

Conflicts of Interest

No conflict of interest was declared.

Abbreviations

AI: artificial intelligence

ASA: American Statistical Association

HTML: hypertext markup language

JSON: JavaScript object notation

LLM: large language model

PRISMA: preferred reporting items for systematic reviews and meta-analyses

UI: user interface

UX: user experience

Reference

1. Mason JB, Mitchell JT. Nutritional surveillance. Bull World Health Organ World Health Organization; 1983;61(5):745.
2. Zhou B, Chen Y, Naumova EN. Revealing Health Disparities with Visual Analytics in the Digital Era: From Individual Attributes to Global Patterns in Nutrition Surveillance. In: Son BWK, editor. Landsc Glob Health Inequity Cham: Springer Nature Switzerland; 2024. p. 129–149. doi: 10.1007/978-3-031-60502-4_10ISBN:978-3-031-60502-4
3. Subramanian M, Wojtusciszyn A, Favre L, Boughorbel S, Shan J, Letaief KB, Pitteloud N, Chouchane L. Precision medicine in the era of artificial intelligence: implications in chronic disease management. J Transl Med 2020 Dec;18(1):472. doi: 10.1186/s12967-020-02658-5
4. Mannar V, Micha R, Allemandi L, Afshin A, Baker P, Battersby J, Bhutta Z, Corvalan C, Di Cesare M, Chen K. 2020 global nutrition report: action on equity to end malnutrition. 89023; 2020. Available from: <https://repository.mdx.ac.uk/item/89022> [accessed Feb 18, 2025]
5. Zhou B, Liang S, Monahan KM, Singh GM, Simpson RB, Reedy J, Zhang J, DeVane A, Cruz MS, Marshak A, Mozaffarian D, Wang D, Semenova I, Montoliu I, Prozorovskaia D, Naumova EN. Food and Nutrition Systems Dashboards: A Systematic Review. Adv Nutr 2022 May 1;13(3):748–757. doi: 10.1093/advances/nmac022
6. Micha R, Coates J, Leclercq C, Charrondiere UR, Mozaffarian D. Global Dietary Surveillance: Data Gaps and Challenges. Food Nutr Bull SAGE Publications Inc; 2018 Jun 1;39(2):175–205. doi: 10.1177/0379572117752986
7. The Food Systems Dashboard is a new tool to inform better food policy | Nature Food. Available from: <https://www.nature.com/articles/s43016-020-0077-y> [accessed Jan 28, 2025]
8. Zhou B, Liang S, Monahan KM, El-Abbadi N, Cruz MS, Chen Y, DeVane A, Reedy J, Zhang J, Semenova I. An open-access data platform: global nutrition and health atlas (gnha). Curr Dev Nutr Oxford University Press; 2022;6(4):nzac031.
9. Béné C, Chege CK, Even B, Hernandez RA, Lundy M, Prager SD, Wiegel J. Why building participatory dashboards is key for sustainable food system transformation. Front Sustain Food

Syst Frontiers Media SA; 2024;8:1405670.

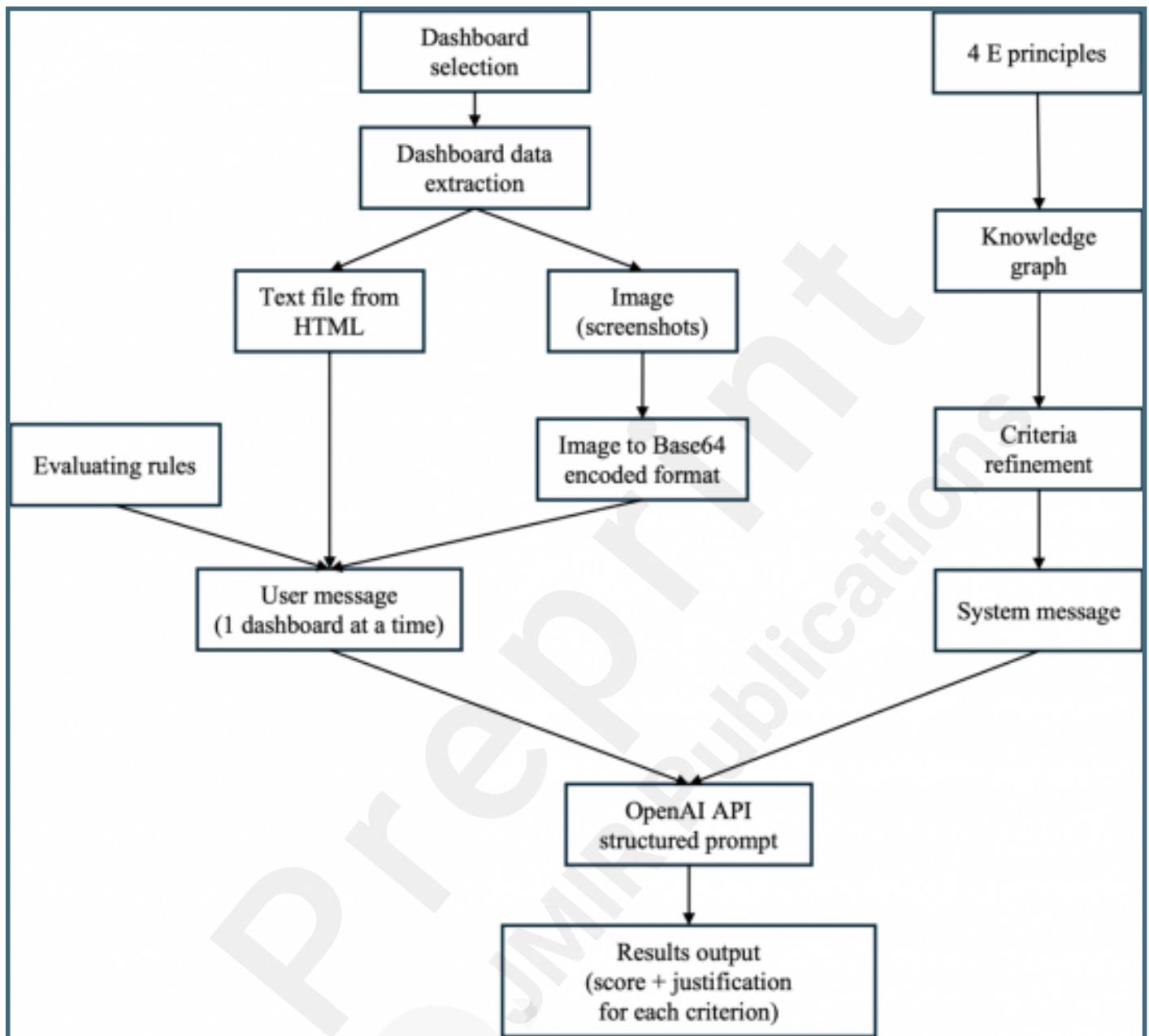
10. Wang L, Zhang S, Wang Y, Lim E-P, Wang Y. LLM4Vis: Explainable Visualization Recommendation using ChatGPT. arXiv; 2023. doi: 10.48550/arXiv.2310.07652
11. Yang Z, Zhou Z, Wang S, Cong X, Han X, Yan Y, Liu Z, Tan Z, Liu P, Yu D, Liu Z, Shi X, Sun M. MatPlotAgent: Method and Evaluation for LLM-Based Agentic Scientific Data Visualization. arXiv; 2024. doi: 10.48550/arXiv.2402.11453
12. Han Y, Zhang C, Chen X, Yang X, Wang Z, Yu G, Fu B, Zhang H. ChartLlama: A Multimodal LLM for Chart Understanding and Generation. arXiv; 2023. doi: 10.48550/arXiv.2311.16483
13. Narechania A, Srinivasan A, Stasko J. NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. IEEE Trans Vis Comput Graph 2021 Feb;27(2):369–379. doi: 10.1109/TVCG.2020.3030378
14. GitHub - holoviz/panel: Panel: The powerful data exploration & web app framework for Python. Available from: <https://github.com/holoviz/panel> [accessed Jan 28, 2025]
15. Sargiotis D. Ethical AI in information technology: Navigating bias, privacy, transparency, and accountability. Adv Mach Lear Art Inte 2024;5(3):01–14.
16. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, López de Prado M, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf Fusion 2023 Nov 1;99:101896. doi: 10.1016/j.inffus.2023.101896
17. Landschaft A, Antweiler D, Mackay S, Kugler S, Rüping S, Wrobel S, Höres T, Allende-Cid H. Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. Int J Med Inf 2024 Sep 1;189:105531. doi: 10.1016/j.ijmedinf.2024.105531
18. Naumova EN. Visual analytics for immunologists: Data compression and fractal distributions. Self/Nonself 2010 Jul;1(3):241–249. PMID:21487480
19. Naumova EN. Telling truth with data visuals: a guide for public health professionals. J Public Health Policy 2024 Jun 1;45(2):191–197. doi: 10.1057/s41271-024-00479-0
20. Naumova EN. Causal AI for public health research and policy: a journey back to the future. J Public Health Policy 2025 Mar 1;46(1):1–7. doi: 10.1057/s41271-024-00541-x
21. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content. J Med Internet Res 2024 Apr 25;26(1):e55847. doi: 10.2196/55847
22. Shi Y, Sha AS, Deng B. Towards Systematic Evaluation of User- Generated Vs AI-Generated Contents in Online Knowledge Exchange Communities. Rochester, NY: Social Science Research Network; 2022. doi: 10.2139/ssrn.4449665
23. Uddin MKS. A Review of Utilizing Natural Language Processing and Ai for Advanced Data Visualization in Real-time Analytics. Rochester, NY: Social Science Research Network; 2024. doi: 10.62304/ijmisds.v1i04.185

24. Wu Z, Chen J, Li Y, Deng Y, Zhao H, Hsieh C-Y, Hou T. From Black Boxes to Actionable Insights: A Perspective on Explainable Artificial Intelligence for Scientific Discovery. *J Chem Inf Model* American Chemical Society; 2023 Dec 25;63(24):7617–7627. doi: 10.1021/acs.jcim.3c01642
25. Lysaght T, Lim HY, Xafis V, Ngiam KY. AI-Assisted Decision-making in Healthcare. *Asian Bioeth Rev* 2019 Sep 1;11(3):299–314. doi: 10.1007/s41649-019-00096-0
26. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges | Vicinagearth. Available from: <https://link.springer.com/article/10.1007/s44336-024-00009-2> [accessed Feb 24, 2025]
27. Naumova EN. A mistake-find exercise: a teacher's tool to engage with information innovations, ChatGPT, and their analogs. *J Public Health Policy* 2023 Jun 1;44(2):173–178. doi: 10.1057/s41271-023-00400-1
28. Naumova EN. Intellectual humility in public health training, research, and practice. *J Public Health Policy* 2023 Mar 1;44(1):1–5. doi: 10.1057/s41271-022-00389-z
29. Naumova EN. Artificial intelligence and data analytics competencies for public health professionals. *J Public Health Policy* 2024 Sep 1;45(3):407–412. doi: 10.1057/s41271-024-00499-w
30. Naumova EN. AI Applications for Nutrition and Food Security Research: A Taxonomy and Competencies. *Nutr Today* 2025 Feb;60(1):20. doi: 10.1097/NT.0000000000000708

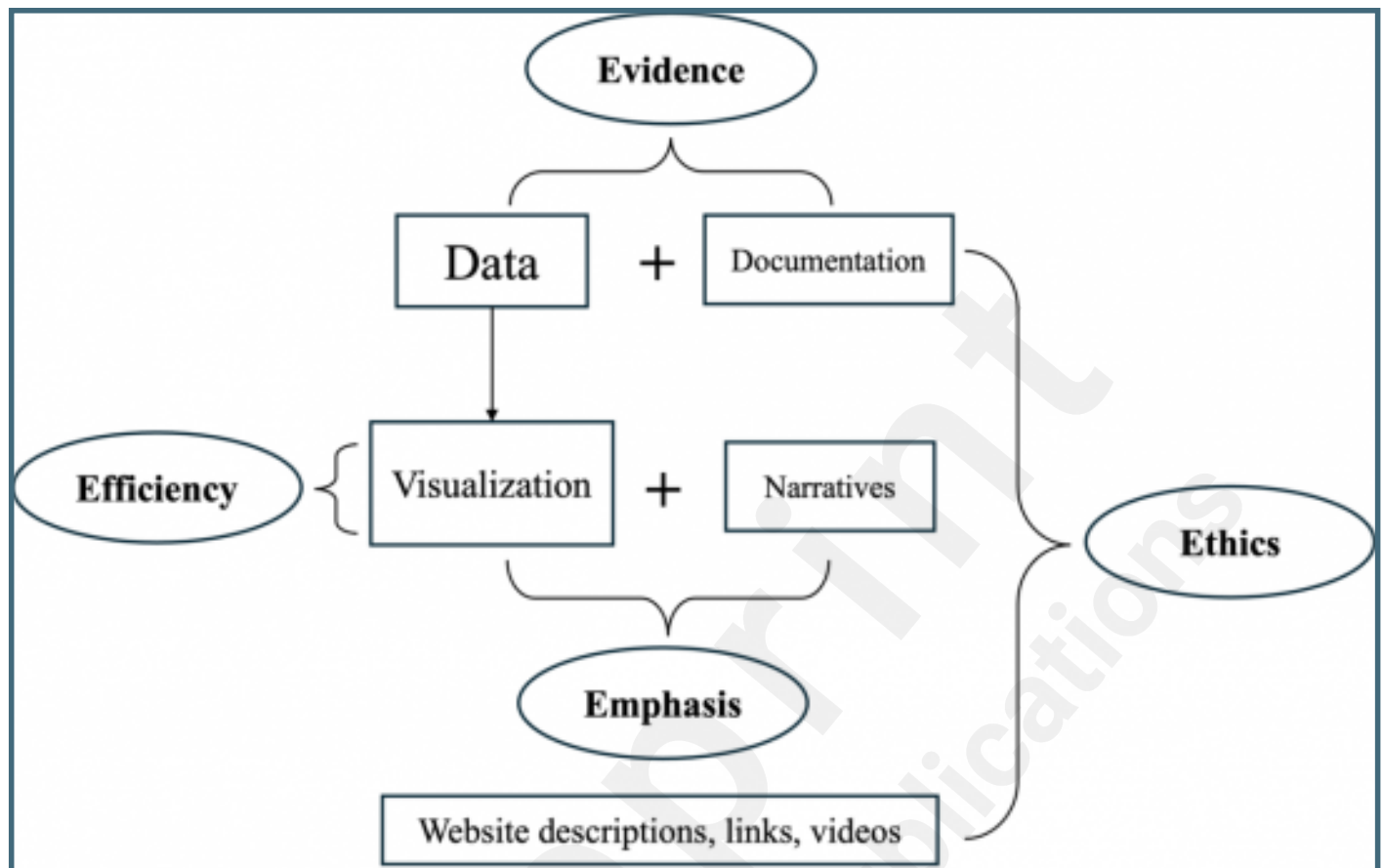
Supplementary Files

Figures

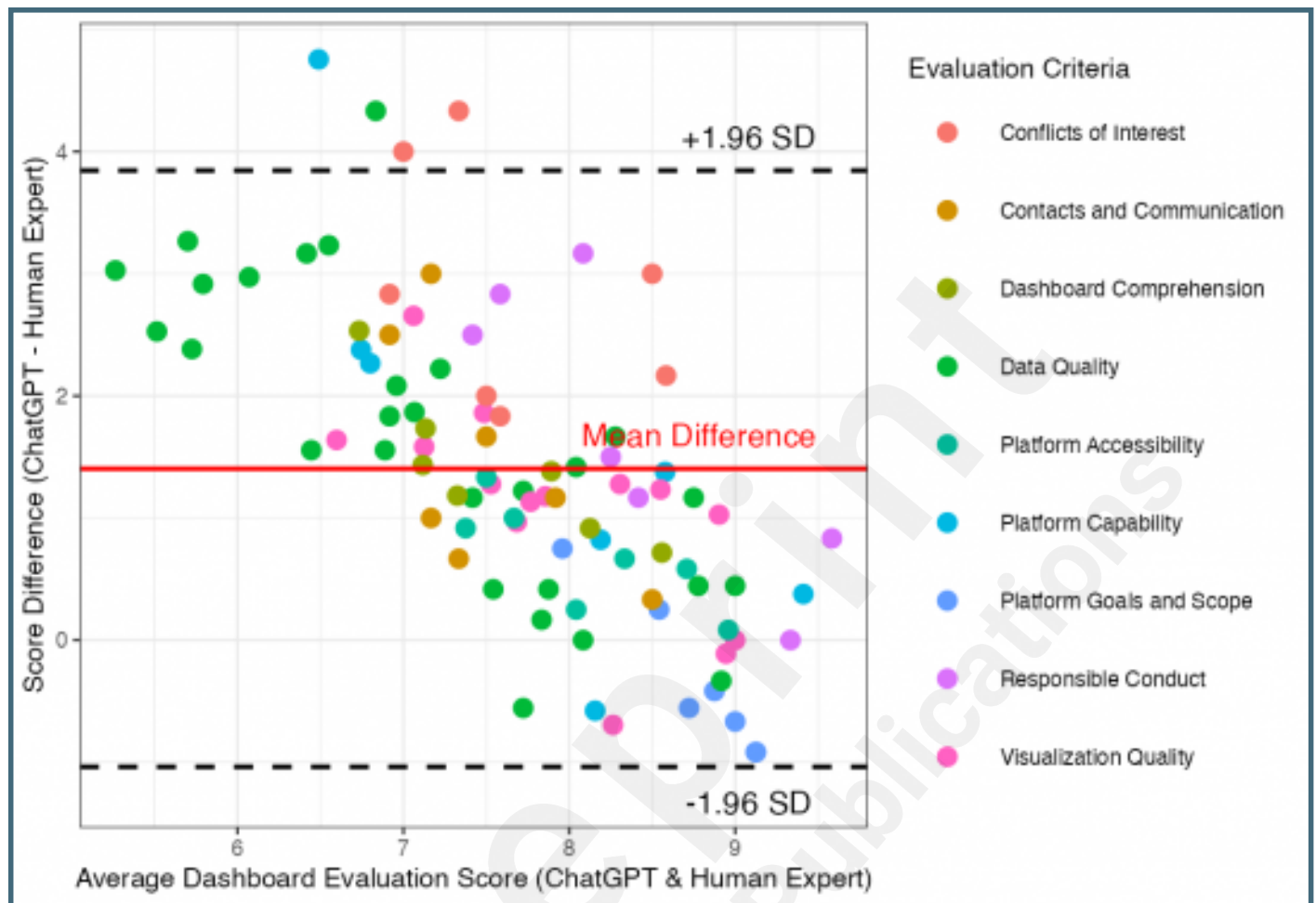
Overview of the dashboard evaluation process.



Knowledge graph framework mapping core 4E principles and dashboard components.



Agreement between ChatGPT-4o and experts in dashboard evaluation.



Multimedia Appendixes

Summary of detailed descriptions of selected dashboards.

URL: <http://asset.jmir.pub/assets/5d0e29febf8b7a3d7a3475c24dd8f43d.xlsx>

Detailed ChatGPT evaluation scores and justifications for all dashboards.

URL: <http://asset.jmir.pub/assets/f7c355b272035581ed9bcce70d54857a.docx>

