https://doi.org/10.31891/2307-5732-2026-361-28
УДК 004.89:004.51

KUZIKOV BORYS
Sumy State University
https://orcid.org/0000-0002-9511-5665
e-mail: b.kuzikov@dl.sumdu.edu.ua
SHOVKOPLIAS SERHII
Sumy State University
https://orcid.org/0000-0003-1837-0213
e-mail: s.shovkoplyas@student.sumdu.edu.ua

AUTOMATED SEMANTIC ALIGNMENT ASSESSMENT FOR WEB ACCESSIBILITY USING LARGE LANGUAGE MODELS

Automated verification of web content accessibility remains an acute problem, as traditional tools are capable of fully testing only approximately 4 out of 50 Web Content Accessibility Guidelines (WCAG) criteria. One of the criteria that is difficult to verify is WCAG 2.5.3, which requires semantic correspondence between the visible text of an element and its accessible name for users of assistive technologies. The objective of this study is to verify the feasibility of using large language models for automated assessment of semantic correspondence according to WCAG 2.5.3 criterion and to determine optimal models by price-quality ratio. A comparative analysis of 17 large language models across different price categories was conducted on a specially created dataset in English and Ukrainian languages. To measure model quality relative to the consensus of leading models, a statistical framework based on consensus assessment was employed with metrics including bias, deviation variance, and coefficient of determination The leading models demonstrated a high level of consistency in semantic similarity assessments (R^2 = 0.85-0.91). Mid-range price segment models showed the best quality-to-cost ratio, notably gemini-2.5-flash-preview achieved the highest consistency (R^2 = 0.91) with minimal noise. The absence of direct correlation between syntactic correctness of responses and the quality of semantic analysis was established. Large language models can be effectively utilized for assessment of WCAG 2.5.3 semantic correspondence. Optimal models for practical application have been identified and directions for further research have been outlined, including knowledge distillation into smaller specialized models to reduce computational costs.

Keywords: *web content accessibility, large language models, semantic similarity, automated testing, assistive technologies*

КУЗІКОВ БОРИС, ШОВКОПЛЯС СЕРГІЙ
Сумський державний університет

АВТОМАТИЗОВАНА ОЦІНКА СЕМАНТИЧНОГО УЗГОДЖЕННЯ У КОНТЕКСТІ ВЕБДОСТУПНОСТІ З ВИКОРИСТАННЯМ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Автоматизована перевірка доступності вебконтенту залишається гострою проблемою, оскільки традиційні інструменти здатні повністю перевірити лише близько 4 із 50 критеріїв стандарту доступності WCAG 2.1. Одним із складних для перевірки є критерій WCAG 2.5.3, який вимагає семантичної відповідності між видимим текстом елемента та його доступним іменем для користувачів асистивних технологій. Метою дослідження є перевірка можливості використання великих мовних моделей для автоматизованої оцінки семантичної відповідності згідно з критерієм WCAG 2.5.3 та визначити оптимальні моделі за співвідношенням ціна-якість. Проведено порівняльний аналіз 17 великих мовних моделей різних цінових категорій на спеціально створеному наборі даних англійською та українською мовами. Використано статистичний фреймворк консенсусної оцінки з метриками зміщення, дисперсії відхилень та коефіцієнта детермінації для вимірювання якості моделей відносно консенсусу провідних моделей. Провідні моделі продемонстрували високий рівень узгодженості в оцінках семантичної схожості (R^2 = 0.85-0.91). Моделі середнього цінового сегменту показали найкраще співвідношення якості та вартості, зокрема gemini-2.5-flash-preview досягла найвищої узгодженості (R^2 = 0.91) з мінімальною шумністю. Встановлено відсутність прямої кореляції між синтаксичною коректністю відповідей та якістю семантичного аналізу. Великі мовні моделі можуть ефективно використовуватися для оцінки семантичної відповідності WCAG 2.5.3. Визначено оптимальні моделі для практичного застосування та окреслено напрямки подальших досліджень, включаючи дистиляцію знань у менші спеціалізовані моделі для зниження обчислювальних витрат.

Ключові слова: *доступність вебконтенту, великі мовні моделі, семантична схожість, автоматизоване тестування, асистивні технології*

Стаття надійшла до редакції / Received 13.11.2025
Прийнята до друку / Accepted 11.01.2026
Опубліковано / Published 29.01.2026

Introduction

Website accessibility is critically important for ensuring equal opportunities of access to information and services for all users, including people with disabilities. Web content must be designed in such a way that it can be perceived and used by people with various visual, auditory, speech, cognitive, and motor impairments. There are sets of rules, such as the WCAG [1], that define accessibility standards which websites must meet.

Existing methods of automated accessibility testing of websites according to WCAG have significant limitations regarding criteria coverage. According to the UsableNet Audit Team, automated tools are capable of fully testing only approximately 4 out of 50 WCAG 2.1 Level A and AA criteria [2]. Although the developers of Axe Core claim coverage of up to 57% of criteria [3], this primarily concerns technical aspects, whereas criteria requiring semantic understanding remain problematic [4].

Automated verification of web content accessibility, specifically WCAG 2.5.3 criterion (Label in Name), remains a complex task [4]. This criterion requires semantic correspondence between the visible text of an element and its "accessible name," which is critical for users of assistive technologies [1]. Traditional algorithms are unable to correctly assess semantic nuances, missing dangerous changes ("Delete" → "Delete All") or erroneously blocking appropriate ones ("Save" → "Save Document") [5].

Recent advances in natural language processing have presented large language models (LLM) as a promising solution for semantic analysis in the context of accessibility. LLMs demonstrate the capability to generate image descriptions [6], detect problems in subtitles and assess label quality[7]. Their ability to consider context allows them to evaluate semantic proximity with high accuracy [8].

However, real-time LLM deployment faces challenges: high computational requirements, high latency [8], risk of vendor lock-in, and unstable provider operation. Figure 1 presents an example of the uptime indicator for the Google Gemini 2.0 Flash Lite model on May 9, 2025, according to Open Router data, which illustrates stability problems even among leading providers.

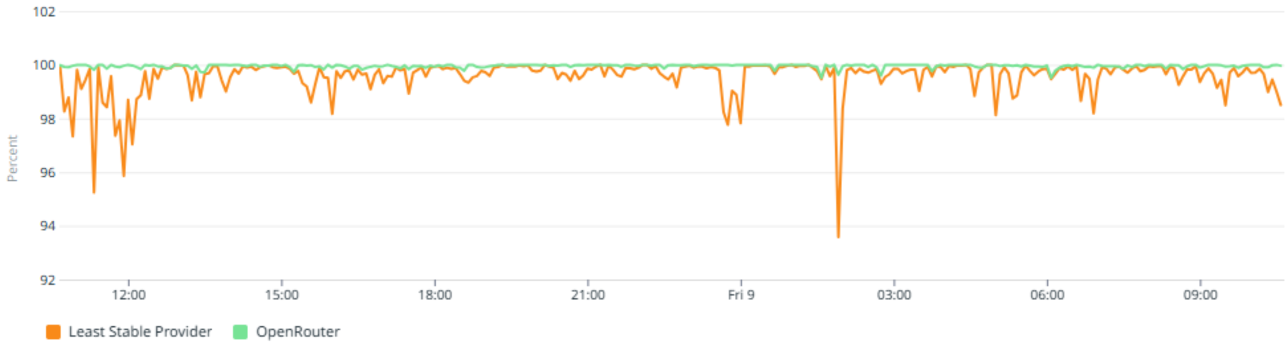

Fig.1. Uptime of Google Model: Gemini 2.0 Flash Lite

The objective of this article is to investigate the possibilities of using large language models for automated assessment of semantic correspondence according to WCAG 2.5.3 criterion, utilizing a specially developed dataset for their verification [9].

Semantic Similarity Analysis Using LLM

To evaluate semantic similarity, LLMs received annotating tasks for "visible text – accessible name" pairs through a specialized prompt framework. The prompt instructed models to apply a scoring scale from -1.0 (semantically opposite or contradictory) to 1.0 (perfect semantic match), enabling clear separation between hazardous semantic divergences and harmless contextual expansions. For cost reduction purposes, input strings were aggregated into 100-pair batches, which substantially degraded results from less capable models. Processing inputs individually in sequential order would have enhanced these models' performance, as batching increases the cognitive load related to contextual analysis and information overload tolerance. Processing incorporated all available pair scores, even when models returned incomplete outputs (including auxiliary comments or partial pair coverage). Temperature was set to 0.1 throughout the experiment to maintain response determinism.

When selecting models among the multitude of available options, we were guided by the following key criteria:

- estimated response quality according to standardized test sets (GPQA, MMLU, DROP);
- number of parameters (which often correlates with quality and performance);
- processing cost;
- latency and time to first token (TTFT).

Processing cost and other operational characteristics largely depend on the model provider. For example, according to Artificial Analysis, for the llama 4 Maverick model, options are available with context windows ranging from 8 thousand to 1 million tokens, processing costs from $0.26 to $0.96 USD per million input tokens, and throughput performance spanning 46 to 721 tokens per second [10].

When conducting the experiment, the context window size was not a decisive factor due to the small volumes of generated results. OpenRouter platform was used as an intermediary for provider selection, which ensures dynamic selection of a provider with optimal latency indicators. In the study, models were grouped by the blended price metric, computed by applying a 3:1 weighting to input versus output token pricing

Table 1 [9] displays model performance data, including the success rate of batch-processed pairs across different test datasets (synthetic EN, UA). This allows assessment of the formal suitability of models for the task, as well as the influence of language on quantitative characteristics of results.

Results demonstrate that most of the mid-range and high-range models can process tasks with high formal success rate (>95%), whereas models with fewer parameters show significantly lower performance, especially for Ukrainian texts.

Table 1

Quantitative capability of models for semantic similarity assessment

Model	Cost, $[3]	Model size, 1×10^9	Success ratio, %	
			EN	UA
Processing cost[4] [0.92$...3.5$]				
gpt-4.1	2.00/8.00	300	**100**	**100**
claude-3.5-haiku	0.80/4.00	~10-20[1]	85.1	91.9
deepseek-prover-v2	0.50/2.18	671	98.8	**100**
deepseek-r1	0.50/2.18	671/ 37[2]	**100**	**100**
Model	**Cost, $[3]**	**Model size,** 1×10^9	**EN**	**UA**
Processing cost[4] (0.1$...0.3$)				
llama-4-maverick	0.17/0.60	400/17[2]	**100**	99.3
gemini-2.5-flash-preview	0.15/0.60	80	**100**	99.8
gpt-4o-mini	0.15/0.6	~10-50[1]	98.9	99.3
qwen3-235b-a22b	0.15/0.60	235/ 22[2]	99.8	99.3
gemini-2.0-flash-001	0.10/0.40	~10-50[1]	**100**	**100**
gpt-4.1-nano	0.10/0.40	~<10[1]	**100**	98.5
Processing cost[4] [0.01$...0.1$]				
ministral-8b	0.10/0.10	8	89.2	90.9
nova-micro-v1	0.035/0.14	~ <10[1]	99.4	99.8
llama-3.1-8b-instruct	0.02/0.03	8	90.9	86.9
llama-3.2-3b-instruct	0.01/0.02	3	65.2	39.5
lfm-3b	0.02/0.02	3	47.2	10.3
qwen2.5-coder-7b-instruct	0.01/0.03	7	84.6	93.9
lfm-7b	0.01/0.01	7	52.7	34.1

[1] Estimated, actual number was not published.

[2] Model built with the Mixture of Experts approach; total number of parameters and number of active parameters are indicated.

[3] Price according to OpenRouter per million input/generated tokens.

[4] Blended processing cost based on 3:1 input-to-output token ratio.

Quality Metrics of Models

Below is an analysis of quality metrics of models with a focus on practical applicability for web accessibility tasks. For statistical justification of metrics and validation of the consensus framework, see. A three-model consensus core comprising gpt-4.1, claude-3.5-haiku, and deepseek-r1 served as the baseline for evaluating the relative quality of remaining models. Table 2 [9] and Figure 2 demonstrate metrics of assessment bias ($\bar{d_j}$), "model noise" (σ_j^2), and consistency with consensus (R^2).

Table 2

Quality metrics of models relative to the consensus core

	R^2	$\overline{d_j}$	σ_j^2
Processing cost[4] [0.92$...3.5$]			
gpt-4.1	0,85	0,04	0,05
claude-3.5-haiku	0,90	0,04	0,04
deepseek-prover-v2	0,62	0,00	0,10
deepseek-r1	0,87	0,00	0,05
Processing cost[4] (0.1$...0.3$)			
llama-4-maverick	0,75	-0,03	0,11
gemini-2.5-flash-preview	0,91	-0,05	0,03
gpt-4o-mini	0,71	-0,06	0,10
qwen3-235b-a22b	0,78	0,01	0,09
gemini-2.0-flash-001	0,70	-0,08	0,11
gpt-4.1-nano	-0,69	0,15	0,21
Processing cost[4] [0.01$...0.1$]			
ministral-8b	-0,75	0,28	0,27

Table 2

	R^2	$\overline{d_j}$	σ_j^2
nova-micro-v1	0,37	0,02	0,22
llama-3.1-8b-instruct	0,18	0,05	0,29
llama-3.2-3b-instruct	0,22	-0,07	0,41
lfm-3b	0,63	0,17	0,35
qwen2.5-coder-7b-instruct	0,07	0,16	0,29
lfm-7b	0,76	0,11	0,18

Results demonstrate a clear dependency between quality metrics of models and their characteristics:

1. High-quality models (premium segment, price >$0.5):
   - claude-3.5-haiku shows the best consistency ($R^2 = 0.90$) with minimal noise ($\sigma^2 = 0.04$);
   - deepseek-r1 demonstrates zero bias and high consistency ($R^2 = 0.87$);
   - gpt-4.1 has stable performance ($R^2 = 0.85$) with low noise.
2. Mid-range price segment ($0.1–$0.3):
   - gemini-2.5-flash-preview achieves the highest consistency among all models ($R^2 = 0.91$) with minimal noise ($\sigma^2 = 0.03$), making it the most effective in terms of price-to-quality ratio;
   - most models show moderate consistency ($R^2 = 0.70–0.78$);
   - gpt-4.1-nano demonstrates negative consistency ($R^2 = -0.69$), indicating systematic discrepancies with consensus.
3. Budget segment (<$0.1):
   - most models shows weak consistency ($R^2 < 0.40$);
   - lfm-7b stands out among budget models with $R^2 = 0.76$, but with elevated bias;
   - ministral-8b has the worst performance ($R^2 = -0.75$) with high noise and significant bias.

Among other findings, the metrics demonstrate the absence of a direct relationship between syntactic correctness and the quality of semantic analysis in models. Thus, the nova-micro-v1 model with high formal success rate (99.4% EN) shows low semantic consistency ($R^2 = 0.37$). More powerful models (>100B parameters) typically demonstrate better consistency and lower noise. Exceptions exist – some smaller specialized models (lfm-7b) show results comparable to much larger models. However, practical application faces low operational stability indicators – 52.7% for the English-language dataset, 34.1% for the Ukrainian-language dataset (see Table 1).

Overall, the results confirm that for the task of semantic assessment according to WCAG 2.5.3, mid-range price segment models are optimal, combining high consistency with acceptable processing cost.

Conclusions

The study demonstrated that large language models can be effectively utilized to solve the task of assessing semantic similarity between visible text labels and their accessible names in the context of ensuring web interface accessibility, specifically for WCAG 2.5.3 criterion.

The -1.0 to 1.0 scoring scale enabled LLMs not only to quantify similarity levels but also to identify semantically contradictory texts, which is critical for detecting accessibility violations and potential malicious attacks (Accessibility Cloaking Attacks [5]).

Leading LLMs demonstrated a high level of consistency in semantic similarity assessments: gemini-2.5-flash-preview ($R^2 = 0.91$), claude-3.5-haiku ($R^2 = 0.90$), deepseek-r1 ($R^2 = 0.87$). This validates the LLM-as-a-Judge methodology as a dependable approach for annotating data when ground truth labels are unavailable. Mid-range price segment models ($0.1–$0.3 per million tokens) showed the best quality-to-cost ratio, making them practically applicable for assessing large volumes of data.

Results demonstrated that models' capacity to produce syntactically correct responses does not directly correlate with their semantic analysis performance, underscoring the requirement for comprehensive evaluation methodologies.

The obtained results confirm the feasibility of using LLMs for automated assessment of WCAG 2.5.3 semantic correspondence, identify optimal models by price-to-quality ratio for practical application, and establish a foundation for further development of effective accessibility testing tools.

Despite positive results, it is important to consider limitations associated with using LLMs in real-world accessibility testing systems:

1. Even budget models require significant resources for processing large volumes of data.
2. For large projects, the cost of API calls can be substantial (from $0.01 to $8.00 per million tokens).
3. As shown in Figure 1 in the Introduction, even leading providers do not always ensure stable operational performance.
4. Dependence on external APIs (vendor lock-in) creates risks for long-term system support.
5. Processing time can be critical for integration into CI/CD processes.

These factors constitute significant barriers to widespread implementation of LLMs in real-time accessibility testing systems. The results emphasize the necessity of seeking more resource-efficient solutions, such as knowledge distillation into smaller specialized models.

References.

1. Web Content Accessibility Guidelines (WCAG) 2.1 [Electronic resource]. – URL: https://www.w3.org/TR/WCAG21/ (accessed: 10.11.2025).

2. Automated WCAG testing is not enough for web accessibility ADA compliance [Electronic resource]. – 2018. – URL: https://blog.usablenet.com/automated-wcag-testing-is-not-enough-for-web-accessibility-ada-compliance (accessed: 10.11.2025).

3. The Automated Accessibility Coverage Report [Electronic resource]. – URL: https://www.deque.com/automated-accessibility-testing-coverage/ (accessed: 10.11.2025).

4. Sane P. A brief survey of current software engineering practices in continuous integration and automated accessibility testing / P. Sane // 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). – 2021. – P. 130–134. DOI: 10.1109/WISPNET51692.2021.9419464.

5. Kuzikov B. Detection and prevention of accessibility cloaking attacks / B. Kuzikov, Tytov P., Shovkoplias O., Lavryk T., Koval V., Kuzikova S. // Information Technology: Computer Science, Software Engineering and Cyber Security. – 2025. – № 1. – P. 124–135. DOI: 10.32782/IT/2025-1-17

6. Improving accessibility through leveraging large language models (LLMs) [Electronic resource]. – 2025. – URL: https://www.deque.com/axe-con/sessions/improving-accessibility-through-leveraging-large-language-models-llms/ (accessed: 10.11.2025).

7. Fatiul Huq S. Automated generation of accessibility test reports from recorded user transcripts / S. Fatiul Huq, M. Tafreshipour, K. Kalcevich, S. Malek // IEEE/ACM 47th International Conference on Software Engineering (ICSE), Apr 26 – May 06 2025. Ottawa, Ontario, Canada. – 2025. - P. 204-216. DOI: 10.1109/ICSE55347.2025.00043

8. Delnevo G. On the interaction with large language models for web accessibility: implications and challenges / G. Delnevo, M. Andruccioli, S. Mirri // 2024 IEEE 21st Consumer Communications &amp; Networking Conference (CCNC): proceedings, 6-9 January 2024. Las Vegas, NV, USA. – 2024. – P. 1–6. DOI: 10.1109/CCNC51664.2024.10454680

9. Kuzikov B., Shovkoplias O., Tytov P., Shovkoplias S., Shutylieva O., Vlasenko O. A Statistical Framework for Consensus-Based Reliability Assessment in Large Language Model Evaluation Applied to Web Accessibility. 5th International Conference on Machine Learning and Big Data Analytics (ICMLBDA) 2025. ICS Global, November 7-8, 2025. DOI: 10.21203/rs.3.rs-8093408/v1

10. Llama 4 Maverick - Intelligence, performance & price analysis: Artificial analysis Model & API Providers Analysis. [Electronic resource]. – URL: https://artificialanalysis.ai/models/llama-4-maverick (Accessed: 10.11.2025).