UDC 004.89:004.51 DOI https://doi.org/10.35546/kntu2078-4481.2025.4.2.12

B. O. KUZIKOV
Candidate of Technical Sciences, Associate Professor,
Associate Professor at the Department of Computer Science
Sumy State University
ORCID: 0000-0002-9511-5665

O. V. VLASENKO
Candidate of Physical and Mathematical Sciences,
Assistant at the Department of Computer Science
Sumy State University
ORCID: 0000-0003-4315-5654

O. V. SHUTYLIEVA
Candidate of Physical and Mathematical Sciences,
Assistant at the Department of Computer
Science Sumy State University
ORCID: 0000-0002-7236-8555

O. A. SHOVKOPLYAS
Candidate of Physical and Mathematical Sciences, Associate Professor,
Head of the Department of Computer
Science Sumy State University
ORCID: 0000-0002-4596-2524

S. R. SHOVKOPLIAS
Postgraduate Student at the Department of Computer Science
Sumy State University
ORCID: 0000-0003-1837-0213

P. O. TYTOV
Postgraduate Student at the Department of Computer Science
Sumy State University
ORCID: 0009-0003-6911-5463

MULTI-MODEL APPROACH TO SYNTHETIC DATA GENERATION FOR SEMANTIC SIMILARITY ASSESSMENT OF WEB INTERFACE TEXT LABELS

Automating web accessibility verification against WCAG standards remains a significant challenge, particularly for criteria requiring semantic understanding of content. WCAG Success Criterion 2.5.3 requires consistency between the visible text of interface elements and their accessible names; however, traditional string matching methods fail to account for semantic nuances of textual modifications. ***Objective.*** *To develop a methodology for creating and validating a robust dataset for assessing semantic correspondence of text labels in the absence of an objective ground truth. This includes developing a taxonomy of semantic transformations, applying a multi-model approach to generation and annotation of synthetic data in Ukrainian and English languages.* ***Methods.*** *The study employs a multi-model approach involving four leading LLMs to generate over 14 thousand unique text pairs and 17 models for data annotation. A taxonomy of semantic transformations was developed that classifies modification types from permissible contextual extensions to critical contradictions. For validation, a statistical framework for consensus formation based on the median scores of a reference core of models was applied, utilizing ICC2k metrics, coefficients of determination, and Spearman correlation.* ***Results****. A validated dataset was created with synthetic data annotated on a semantic similarity scale from -1 to 1. The multi-model approach ensured dataset diversity and minimized biases of individual models.* ***Conclusions.*** *The developed methodology effectively addresses the problem of creating training data without an objective ground truth. The formed dataset enables objective comparison of commercial LLMs and cost-effective knowledge distillation into small models for practical application in automated accessibility testing tools.*

Key words: *large language models, web accessibility, multi-model approach, semantic similarity, statistical consensus, synthetic training data, dataset validation.*

Б. О. КУЗІКОВ
кандидат технічних наук, доцент,
доцент кафедри комп'ютерних наук
Сумський державний університет
ORCID: 0000-0002-9511-5665

О. В. ВЛАСЕНКО
кандидат фізико-математичних наук,
асистент кафедри комп'ютерних наук
Сумський державний університет
ORCID: 0000-0003-4315-5654

О. В. ШУТИЛЄВА
кандидат фізико-математичних наук,
асистент кафедри комп'ютерних наук
Сумський державний університет
ORCID: 0000-0002-7236-8555

О. А. ШОВКОПЛЯС
кандидат фізико-математичних наук, доцент,
завідувач кафедри комп'ютерних наук
Сумський державний університет
ORCID: 0000-0002-4596-2524

С. Р. ШОВКОПЛЯС
аспірант кафедри комп'ютерних наук
Сумський державний університет
ORCID: 0000-0003-1837-0213

П. О. ТИТОВ
аспірант кафедри комп'ютерних наук
Сумський державний університет
ORCID: 0009-0003-6911-5463

МУЛЬТИМОДЕЛЬНИЙ ПІДХІД ДО ГЕНЕРАЦІЇ СИНТЕТИЧНИХ ДАНИХ ДЛЯ ОЦІНКИ СЕМАНТИЧНОЇ СХОЖОСТІ ТЕКСТОВИХ МІТОК ВЕБІНТЕРФЕЙСІВ

Автоматизація перевірки веб-доступності за стандартами Web Content Accessibility Guidelines (WCAG) залишається значним викликом, особливо для критеріїв, що вимагають семантичного розуміння контенту. Критерій WCAG 2.5.3 «Мітка у імені» вимагає узгодженості між видимим текстом елементів інтерфейсу та їхніми доступними імена-ми, проте традиційні методи зіставлення рядків не враховують семантичні нюанси текстових модіфікацій. ***Мета.*** *Розробити методологію створення та валідації надійного набору даних для оцінки семантичної відповідності текстових міток за відсутності об'єктивного еталону. Це включає розробку таксономії семантичних змін, застосування мультимодельного підходу до генерації та анотації синтетичних даних українською та англійською мовами.* ***Методи.*** *Дослідження використовує мультимодельний підхід із залученням чотирьох провідних LLM для генерації понад 14 тисяч унікальних пар текстів та 17 моделей для анотування даних. Розроблено таксономію семантичних змін, що класифікує типи модіфікацій від допустимих розширень контексту до критичних суперечностей. Для валідації застосовано статистичний фреймворк формування консенсусу на основі медіани оцінок еталонного ядра моделей із використанням метрик ICC2k, коефіцієнтів детермінації та кореляції Спірмена.* ***Результати.*** *Створено валідований датасет, синтетичні данні анотовані за шкалою семантичної схожості від -1 до 1. Мультимодельний підхід забезпечив різноманітність датасету та мінімізував упередження окремих моделей.* ***Висновки.*** *Розроблена методологія ефективно вирішує проблему створення навчальних даних за відсутності еталону. Сформований датасет уможливлює об'єктивне порівняння комерційних LLM та економічно ефективну дистиляцію знань у малі моделі для практичного застосування в інструментах автоматизованого тестування доступності.*

Ключові слова: *великі мовні моделі, доступність вебсайтів, мультимодельний підхід, семантична схожість, статистичний консенсус, синтетичні навчальні дані, валідація наборів даних.*

Introduction

Website accessibility ensures equitable information and service access for all users, particularly individuals with disabilities, by accommodating diverse impairments affecting vision, hearing, speech, cognitive function, and motor abilities. The Web Content Accessibility Guidelines (WCAG) establish accessibility standards that websites must meet.

Automating WCAG compliance verification presents considerable challenges. Contemporary tools can automatically assess only a limited subset of criteria [1], focusing predominantly on technical accessibility aspects. However, criteria demanding semantic content comprehension remain resistant to automation [2]. WCAG Success Criterion 2.5.3 (Label in Name) mandates that the visible text of an interface element must be contained within its accessible name, ensuring alignment between what users perceive visually and what assistive technologies interpret [3]. Current automated accessibility testing tools rely primarily on rudimentary string-matching algorithms, such as substring inclusion or Levenshtein distance. These approaches fail to capture semantic nuances in textual modifications: they may flag legitimate contextual additions ("Process" → "Process payment") as violations while simultaneously overlooking critical semantic alterations ("Save" → "Seve and delete").

Large language models (LLMs) demonstrate capability in accounting for semantic relationships and contextual understanding [4-6]. Nevertheless, their application to training data generation encounters a fundamental challenge: the absence of objective ground truth. Reliance on a single LLM carries the risk of inheriting its systematic biases.

The objective of this work is to develop a methodology for creating and validating a robust dataset for assessing semantic correspondence of text labels according to WCAG Success Criterion 2.5.3 in the absence of an objective ground truth. This encompasses: (1) development of a taxonomy of semantic transformations to systematize relationships between texts, (2) application of a multi-model approach to synthetic data generation to avoid biases of individual models, and (3) creation of a statistical framework for forming consensus-based scores as "proxy ground truth" based on aggregation of results from multiple LLMs.

Dataset Creation and Validation

To systematize differences between visible text and its ARIA description, a taxonomy was developed that classifies various types of semantic modifications and their potential impact on accessibility. The taxonomy serves as a fundamental foundation for subsequent synthetic dataset generation, as it enables purposeful creation of examples covering the entire spectrum of possible semantic relationships between texts.

Categories are ordered by increasing potential risk–from permissible modifications that enhance accessibility to potentially hazardous changes that may mislead users. The developed taxonomy is presented in Table 1.

Table 1

Taxonomy of Semantic Transformations

Category	Subcategory	Visible Text	ARIA-label	Criticality
Contextual Extension	Addition of Clarification	Download	Download user manual	Low
	Addition of System Information	Download	Download (Ctrl+D)	Low
Action Object Modification	Direct Object Modification	Process payment	Process order	Medium
	Object Addition	Download	Download and delete file	Medium
Action Type Modification	Addition of Opposite Action	Save	Save and delete	High
	Complete Action Change	Download	Cancel download	High
Negation	Direct Negation	Download	Do not download	Critical
	Contextual Negation	Download	Disable download function	Critical
Technical Modifications	Formatting	Download file	DOWNLOAD FILE	Low
	Abbreviations	Download frequently asked questions	Download (FAQ)	Low

This classification provides a structured approach to test data generation covering all relevant scenarios from safe contextual extensions to critical semantic contradictions.

Multi-Model Synthetic Data Generation

To avoid biases of individual models and ensure maximum data diversity, synthetic dataset generation was conducted using four leading commercial LLMs: Anthropic Claude Sonnet (versions 3.5 and 3.7), OpenAI ChatGPT 4o, Google Gemini (2.0 Experimental Advanced, 2.5 Pro Experimental), and Grok 3.

When formulating queries to the models, two key requirements were established reflecting the specificity of WCAG Success Criterion 2.5.3:

1. The visible text must be part of the alternative text.
2. The alternative text should preferably lead with the visible text.

The prompt for data generation was structured as follows:

```
I need you to generate examples of web accessibility text alternatives that follow a
specific pattern of {EXPLANATIONS}
Requirements:
```

```
Generate 100 examples in {LANGUAGE}
Each example should have:
Visible text: short, clear action or label. Mostly, more than one word.
Alternative text: expanded version that {EXPLANATIONS}

Format: "visible_text[TAB]alternative_text"
The visible text MUST be included within the alternative text

Preferably place the visible text at the beginning of the alternative text.
Examples should be **realistic** and represent common web interface elements - label for
input field OR button.

Example format:
Submit[TAB]Submit registration form
Send[TAB]Send message to support team
Download[TAB]Download user manual in PDF format
The examples should focus on clarification additions, NOT technical/system information.
For instance:

Good: "Delete[TAB]Delete selected items permanently"
Not: "Delete[TAB]Delete (Ctrl+D)"

The examples should represent realistic web interface scenarios like:

Form actions
Navigation elements
Content interactions
User account operations
Media controls
Document operations
Communication features

Please generate 100 such examples, each on a new line, maintaining consistent formatting
and ensuring real-world applicability.
```

The {EXPLANATIONS} parameter was substituted with the description of the corresponding category from the taxonomy (Table 1), while {LANGUAGE} was replaced with "English" or "Ukrainian" to create English-language and Ukrainian-language datasets respectively.

The total volume of generated data comprised over 14,384 unique text pairs:

- ChatGPT 4o: 3 429 samples;
- Claude Sonnet (3.5/3.7): 4 004 samples;
- Google Gemini (2.0/2.5): 4 470 samples;
- Grok 3: 2 481 samples.

Each taxonomy category was represented by 200–400 examples from each model. The utilization of multiple models not only increased data volume but also ensured diversity of stylistic and lexical characteristics. For instance, only ChatGPT employed emoji symbols in generated responses, demonstrating unique characteristics of each model and confirming the validity of the multi-model approach.

LLM-Based Annotation of Generated Data

Following the generation of synthetic text pairs, the critical task of their annotation arose–assigning each pair a numerical semantic similarity score. Since the scale of data (over 14 000 pairs) renders manual annotation unrealistic, and reliance on a single LLM carries the risk of inheriting its biases, a multi-model annotation approach with subsequent statistical score consensus was applied.

For semantic similarity assessment, a specialized prompt was developed that instructs models to evaluate pairs on a scale from -1.0 to 1.0, where -1.0 represents opposite or contradictory meaning and 1.0 represents perfect semantic correspondence. Few-shot prompting technique was employed in prompt development – providing the model with several examples of input data together with the expected output format directly in the query text. This approach enables the model to better comprehend task specificity without additional training. The data annotation prompt is presented below.

```
You are a semantic similarity analyzer for web accessibility. Analyze the semantic similarity between visible text and ARIA labels according to these rules:

**Similarity Scale**:
1.0: Perfect semantic match (identical meaning)
0.8 to 0.95: Context extension without meaning change
0 for completely unrelated text
0.4 to 0.7: Object or scope variation, core action preserved
-0.2 to 0.4: Different but related actions
-1.0 to -0.2: Opposite or contradictory meaning

**Core Rules**:
1. Symmetrical comparison (order doesn't matter)
2. Identical texts always get 1.0
3. Direct opposites always get -1.0
4. Unrelated texts always get 0
5. Technical annotations have minimal impact on score
6. Do not add numbering.
7. Just answer, **no explanation**.
8. Do not ask question, always process all input.

Input Format:
The input is provided as two texts separated by a tab character:
<text1>[TAB]<text2>

Output Format:
<text1>[TAB]<text2>[TAB]<similarity_score>

**Examples**:
Input: Submit[TAB]Submit form
Output:
Submit[TAB]Submit form[TAB]0.85,

Input: Submit[TAB]Do not submit
Output:
Submit[TAB]Do not submit[TAB]-1.0

**Compare the following pair**:
```

To conserve computational resources and reduce costs, input pairs were batched in groups of 100 samples per query. Such batch processing imposes greater demands on mdel resilience and contextual analysis to information overload, which may adversely affect the quality of weaker models. A temperature of 0.1 was used to ensure maximum response determinism. During processing, all available scores were considered, even when a model returned partial results.

Statistical Framework for Consensus Formation

Quality assessment of large dataset annotation traditionally requires substantial human effort. To address this challenge, the LLM-as-a-Judge approach [7] was applied, wherein large language models serve as automated evaluators. In this study, 17 different LLMs were engaged to annotate each text pair. Given the absence of an objective "correct" value (ground truth), model reliability was assessed relative to a consensus score formed by a "core" of the most reliable models.

Key Framework Principles:

1. A set of reference models is determined (e.g., the most powerful LLMs with proven quality). The consensus score for each sample is obtained by taking the median of core scores. The median was selected due to its robustness to outliers and systematic shifts.
2. For each model, deviation metrics from consensus are computed (systematic model biases), including deviation variance (characterizing model "noisiness") and coefficient of determination R^2 (indicating concordance with the reference core).
3. An approach is employed that balances consensus quality and computational costs, enabling formation of a minimal model core while maintaining high reliability.

A detailed mathematical description of the framework, including formal metric definitions, theoretical properties (median robustness, ICC monotonicity), and optimization algorithms is presented in the accompanying publication [8]. Application of these metrics enables objective model comparison even no objective ground truth is available, focusing on their internal consistency. The formed consensus serves as "proxy ground truth" for annotation of the generated dataset and subsequent training of small language models.

Conclusions

The developed methodology for dataset creation and validation for semantic similarity assessment of text labels demonstrates the effectiveness of a multi-model approach under conditions of objective ground truth absence.

The developed taxonomy of semantic transformations (Table 1) provided a systematic approach to synthetic data generation, encompassing the entire spectrum of possible relationships between visible text and ARIA description– from safe contextual extensions to critical semantic contradictions. Using a semantic similarity scale ranging from -1 to 1 proved effective for nuanced assessment, enabling both quantification of semantic proximity and detection of contradictory meanings. This proves essential for identifying potentially hazardous inconsistencies, which may indicate Accessibility Cloaking Attacks[9].

The employment of four different LLMs for data generation ensured high dataset diversity. As demonstrated (e.g., emoji usage only by ChatGPT), each model possesses unique stylistic characteristics. This confirms that reliance on a single model would lead to systematic bias in the data. Generated dataset with annotation is placed at [10]. Similarly, 17 different models were engaged for data annotation, from which a compact evaluator core was formed, enabling robust consensus formation through aggregation of multiple "opinions" and avoidance of individual model bias influence.

The LLM-as-a-judge approach with consensus formation based on median scores of the reference model core proved effective for creating "proxy ground truth". Detailed statistical analysis of concordance among leading LLMs (including ICC2k computation, analysis of determination coefficients R^2 and Spearman correlation) is presented in. A central finding is that despite differences in underlying architecture and training procedures, leading models demonstrate similar semantic understanding patterns, enabling reliable consensus formation. A key finding is the distinction between formal operability (syntactic correctness of responses) and semantic assessment quality, which do not necessarily correlate. For instance, certain models demonstrated a high percentage of successful response generation but low concordance with consensus, and vice versa (see Table 2 in for details). This underscores the necessity of a statistical validation approach rather than simple counting of successful responses.

The created and validated dataset constitutes a key asset enabling both objective comparison of expensive commercial LLMs and cost-effective distillation of their knowledge into small, fast models for practical application in automated web accessibility testing tools. [11]

Bibliography

1. Automated WCAG Testing Is Not Enough for Web Accessibility & ADA Compliance. UsableNet. URL: https://blog.usablenet.com/automated-wcag-testing-is-not-enough-for-web-accessibility-ada-compliance (date of access: Nov. 12, 2025).

2. Sane P. A Brief Survey of Current Software Engineering Practices in Continuous Integration and Automated Accessibility Testing, *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, 2021, pp. 130-134. DOI: https://doi.org/10.1109/WiSPNET51692.2021.9419464

3. Web Content Accessibility Guidelines (WCAG) 2.1. URL: https://www.w3.org/TR/WCAG21/. (date of access: Nov. 12, 2025).

4. Huq S.F., Tafreshipour M., Kalcevich K., Malek S. Automated Generation of Accessibility Test Reports from Recorded User Transcripts, *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*, Ottawa, ON, Canada, 2025, pp. 204-216. DOI: https://doi.org/10.1109/ICSE55347.2025.00043.

5. Delnevo G., Andruccioli M., Mirri S. On the Interaction with Large Language Models for Web Accessibility: Implications and Challenges, *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, USA, 2024, pp. 1-6. DOI: https://doi.org/10.1109/CCNC51664.2024.10454680.

6. Aralimatti R., Shakhadri S.A.G., Kr K, Angadi K. Fine-Tuning Small Language Models for Domain-Specific AI: An Edge AI Perspective, *Preprints*, Feb. 2025. URL: https://doi.org/10.20944/preprints202502.2128.v1 (date of access Nov. 12, 2025).

7. Gu J., Jiang X., Shi Z., Tan H., Zhai X., Xu C., Li W., Shen Y., Ma S., Liu H., Wang S., Zhang K., Wang Y., Gao W., Ni L., Guo, J. A survey on LLM-as-a-Judge. URL: https://doi.org/10.48550/arXiv.2411.15594 (date of access Nov. 12, 2025).

8. 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. https://doi.org/DOI: 10.21203/rs.3.rs-8093408/v1

9. Kuzikov B., Tytov P., Shovkoplias O., Lavryk T., Koval V., Kuzikova S., Detection And Prevention Of Accessibility Cloaking Attacks, *Information Technology Computer Science Software Engineering and Cyber Security*, № 1, p. 124–135. DOI: https://doi.org/10.32782/it/2025-1-17

10. Tytov P.O. Semantic Language Models for WCAG. URL: https://www.kaggle.com/datasets/tytovpavel/semantic-language-models-for-wcag. (date of access Nov. 12, 2025)

11. Kuzikov B.O., Shovkoplias O.A., Tytov P.O., Shovkoplias S.R. Application of small language models for semantic analysis of web interface accessibility, *Проблеми програмування*, №. 2, С. 77–86, 2025. https://doi.org/10.15407/pp2025.02.077

References

1. UsableNet. (n.d.). *Automated WCAG testing is not enough for web accessibility & ADA compliance*. https://blog.usablenet.com/automated-wcag-testing-is-not-enough-for-web-accessibility-ada-compliance

2. Sane, P. (2021). A brief survey of current software engineering practices in continuous integration and automated accessibility testing. In *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 130–134). IEEE. https://doi.org/10.1109/WiSPNET51692.2021.9419464

3. World Wide Web Consortium. (2025). *Web Content Accessibility Guidelines (WCAG) 2.1*. https://www.w3.org/TR/WCAG21/

4. Huq, S. F., Tafreshipour, M., Kalcevich, K., & Malek, S. (2025). Automated Generation of Accessibility Test Reports from Recorded User Transcripts. *In 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)* (pp. 204–216). https://doi.org/10.1109/icse55347.2025.00043

5. Delnevo, G., Andruccioli, M., & Mirri, S. (2024). On the Interaction with Large Language Models for Web Accessibility: Implications and Challenges. *In 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)* (pp. 1–6). IEEE. https://doi.org/10.1109/ccnc51664.2024.10454680

6. Aralimatti, R., Shakhadri, S. A. G., KR, K., & Angadi, K. (2025). Fine-Tuning Small Language Models for Domain-Specific AI: An Edge AI Perspective. MDPI AG. https://doi.org/10.20944/preprints202502.2128.v1

7. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2024). A Survey on LLM-as-a-Judge (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2411.15594

8. Kuzikov, B., Shovkoplias, O., Tytov, P., Shovkoplias, S., Shutylieva, O., & Vlasenko, O. (2025). A statistical framework for consensus-based reliability assessment in large language model evaluation applied to web accessibility. In *5th International Conference on Machine Learning and Big Data Analytics (ICMLBDA) 2025*. ICS Global. https://doi.org/10.21203/rs.3.rs-8093408/v1

9. Kuzikov, B., Tytov, P., Shovkoplias, O., Lavryk, T., Koval, V., & Kuzikova, S. (2025). Detection and prevention of accessibility cloaking attacks. *Information Technology: Computer Science, Software Engineering and Cyber Security*, 1, 124–135. https://doi.org/10.32782/it/2025-1-17

10. Tytov, P. O. (2025). *Semantic language models for WCAG* [Dataset]. Kaggle. https://www.kaggle.com/datasets/tytovpavel/semantic-language-models-for-wcag

11. Kuzikov, B. O., Shovkoplias, O. A., Tytov, P. O., & Shovkoplias, S. R. (2025). Application of small language models for semantic analysis of web interface accessibility. *Problems in Programming*, *2*, 77–86. https://doi.org/10.15407/pp2025.02.077

Дата першого надходження рукопису до видання: 16.11.2025
Дата прийнятого до друку рукопису після рецензування: 12.12.2025
Дата публікації: 31.12.2025