ARWI: Arabic Write and Improve

Kirill Chirkunov,[1] Bashar Alhafni,[1,2] Chatrine Qwaider,[1]
Nizar Habash,[1,2] Ted Briscoe[1]
[1]MBZUAI, [2]New York University Abu Dhabi
{kirill.chirkunov,chatrine.qwaider,ted.briscoe}@mbzuai.ac.ae
{alhafni,nizar.habash}@nyu.edu

Abstract

Although Arabic is spoken by over 400 million people, advanced Arabic writing assistance tools remain limited. To address this gap, we present ARWI, a new writing assistant that helps learners improve essay writing in Modern Standard Arabic. ARWI is the first publicly available[1] Arabic writing assistant to include a prompt database for different proficiency levels, an Arabic text editor, state-of-the-art grammatical error detection and correction, and automated essay scoring aligned with the Common European Framework of Reference standards for language attainment. Moreover, ARWI can be used to gather a growing auto-annotated corpus, facilitating further research on Arabic grammar correction and essay scoring, as well as profiling patterns of errors made by native speakers and non-native learners. A preliminary user study shows that ARWI provides actionable feedback, helping learners identify grammatical gaps, assess language proficiency, and guide improvement.

1 Introduction

Arabic is the national language of over 400 million people and one of the UN's six official languages (Ryding and Wilmsen, 2021; United Nations, 2024). Yet, Arabic writing assistance tools remain severely underdeveloped. Unlike English, which has numerous competitive writing assistants and CEFR-benchmarked grading systems (Council of Europe, 2001), Arabic tools are limited to a few commercial error-correction systems with no objective public evaluation. Enhanced writing assistants could benefit millions of Arabic writers and aid corpus collection, advancing Arabic NLP.

The development of Arabic writing assistants faces major challenges, with one of the most significant being the lack of a diverse Arabic corpus that captures the wide range of writing variations, including grammatical errors made by both native speakers and second language learners. Having such a comprehensive corpus would enable the creation of writing assistants that not only provide accurate error detection and correction suggestions but also motivate learners to continuously enhance their Arabic writing skills. Additionally, these assistants would contribute to ongoing data collection while actively supporting users in refining their writing abilities.

In response to these challenges, we introduce **ARWI**, a writing assistant tool specifically designed to help MSA writers improve their essay-writing skills. **ARWI** features an intuitive interface and user experience based on the following core components:

- **Essay Prompt Database**: A library of writing topics across CEFR levels.
- **Arabic Text Editor**: Highlights errors, aids structuring, and supports iterative drafting.
- **Grammar Error Detection & Correction (GED/C)**: Identifies errors (e.g., orthography, morphology) and offers feedback.
- **Automated Essay Scoring (AES)**: Assesses grammar, vocabulary, and errors to estimate CEFR levels (A1-C2).
- **Progress Tracking**: Stores revisions and visualizes improvement.
- **User Profiling**: Allows learners to specify dialect, native language, and proficiency.
- **Auto-Annotated Corpora**: A growing repository of diverse, auto-annotated essay samples.

Section 2 presents related work; and Section 3 presents a description of the ARWI system. We discuss a preliminary user experiment in Section 4, and our conclusions and outlook in Section 5.

[1] https://arwi.mbzuai.ac.ae/

Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025), pages 11–18

2 Related Work

2.1 Existing datasets for writing improvement

Prominent English datasets include the CoNLL-2014 corpus (Ng et al., 2014)–derived from the NUCLE (Dahlmeier et al., 2013) release with approximately 1.2 million words–along with WI-LOCNESS (Bryant et al., 2019; Granger, 1998) which offers 3,000 annotated essays (628K words) grouped by CEFR levels. More recently, the Write & Improve annotated corpus (Nicholls et al., 2024) has provided a large resource of 23,000 annotated essays with detailed CEFR annotations, supporting both Grammatical Error Detection/Correction (GED/C) and Automatic Essay Scoring (AES) tasks. In addition, several English GED/C datasets such as GMEG-Yahoo and GMEG-Wiki (Napoles et al., 2019) extend the scope by covering different business domains as well as formal and informal speech registers. The JFLEG dataset (Napoles et al., 2017) further complements these resources by focusing on fluency as opposed to minimal meaning-preserving edits.

Arabic datasets are limited in both size and diversity. The QALB-2014 corpus (Mohit et al., 2014) contains around 1.2 million words across 21,396 sentences from online commentaries on Al Jazeera articles, each paired with a corrected version to facilitate GED/C research. QALB-2015 (Rozovskaya et al., 2015) adds another layer by offering 622 annotated essay sentences (approximately 140K words) from both native and non-native writers. Complementing these, the ZAEBUC corpus (Habash and Palfreyman, 2022) comprises 214 annotated Arabic essays (about 33.3K words) with CEFR grades, thus addressing both GED/C and AES tasks. However, even combined, these Arabic resources lack the extensive genre, topic and proficiency-level stratification of their English counterparts.

2.2 Arabic Writing Assistance Tools

In contrast to numerous English writing assistants like Write&Improve,[2] Grammarly, and others (Sanz-Valdivieso, 2024), which assess fluency and grammar, Arabic tools (e.g., Sahehly,[3] Qalam[4]) focus on common errors but lack overall writing quality feedback. They show good performance in identifying and correcting common errors, such as Hamza placement or confusion between Ha, Ta, and Ta-Marbuta, but lack the capability to detect and correct more nuanced error types, such as merge/split errors or issues related to the shortening of long vowels, as outlined in comprehensive error taxonomies (Alfaifi and Atwell, 2012; Alfaifi et al., 2013).

2.3 LLMs as Arabic Writing Assistants

The advent of large language models (LLMs) has led to the development of writing assistants based on zero-shot or few-shot prompt engineering (Fitria, 2023; Yancey et al., 2023; Pack et al., 2024; Kim et al., 2024), as seen in multilingual (ChatGPT, Gemini, etc.) and Arabic-centric LLMs (Jais Chat (Sengupta et al., 2023) and Fanar (Team et al., 2025)). Despite their strong baseline performance, these models tend to fall short when compared to specialized systems focused on GED/C and AES (Wu et al., 2023; Alhafni and Habash, 2025).

Recent fine-tuning experiments on English GED/C and AES datasets have yielded promising results, demonstrating that pretrained LLMs can achieve state-of-the-art performance in GEC (Omelianchuk et al., 2024) if used within ensemble models. This observation underscores the potential benefits of creating a rich, diverse corpus of annotated Arabic texts, which would facilitate the fine-tuning of LLMs specifically for MSA writing assistance.

3 System Description

3.1 Overview of ARWI

ARWI functions as a web application, integrating a front-end user interface with a backend of specialized REST API services and data collection infrastructure. The system includes an Arabic text editor with diacritics support, GED/C auto-annotation, AES module, and progress tracking that displays learning trajectories and revision improvements. ARWI delivers personalized, actionable feedback to help users continuously enhance their writing skills. Screenshots of the system are provided in Figure 1 to illustrate ARWI's current UI/UX and typical pattern of use. Figure 2 in Appendix A shows the English version of the interface.

3.2 Core Components

3.2.1 Collection of Essay Prompts

We develop an expandable database of essay prompts to provide targeted writing tasks for all

[2] https://writeandimprove.com/
[3] https://sahehly.com/
[4] https://qalam.ai/

Figure 1: A before-and-after example of using ARWI's Arabic interface. In (a) the text receives a B1 CEFR and a large number of errors marked with red underlining; in (b) the results shows improved writing and is raised to B2 CEFR. See Appendix A for the English version of the interface.

CEFR levels. Each prompt covers a specific topic across various domains, aligning with Arabic cultural sensitivities and supporting both formal and informal genres. ARWI enforces a minimum word count: 50 words for beginners (A1-A2), 100 for intermediate writers (B1-B2), and 200+ for advanced learners (C1-C2).

Beginner prompts focus on descriptive writing (e.g., favorite animals, family traditions). Intermediate learners engage with reflective or argumentative topics (e.g., pros and cons of wearing uniforms), while advanced writers tackle analytical discussions (e.g., AI ethics, environmental sustainability). Additionally, some prompts include optional media elements, such as images, to support descriptive tasks involving processes, interior spaces, or graphical representations.

Many Arabic proficiency exams, including CIMA[5] and ALPT[6], require writing tasks. Our essay prompt design draws inspiration from these exams, aligning with their task types. By mapping prompts to the CEFR scale, we ensure appropriate difficulty levels and help learners prepare for CEFR-benchmarked Arabic proficiency tests.

3.2.2 Arabic Text Editor

The editor disables *real-time* spell-checking and auto-corrections, instead providing actionable feedback from the GED/C module upon submission. This approach encourages users to review and apply changes manually, reinforcing learning and improving retention. See Figure 1.

3.2.3 GED/C Module

For GED, we adopt a two-stage token-level classification approach, similar to (Alhafni et al., 2023), by fine-tuning CAMeLBERT-MSA (Inoue et al., 2021). The first classifier performs binary GED, identifying whether a token is erroneous, while the second classifier provides a more fine-grained analysis, categorizing errors based on the ARETA taxonomy (Belkebir and Habash, 2021). These classifiers are applied sequentially: the binary classifier runs first, followed by the fine-grained classifier. This cascaded setup ensures high precision in our GED module.

For GEC, we develop a text-editing system that predicts character-level edits for each input token, generating the corrected text when applied (Alhafni and Habash, 2025). Both GED/C models are fine-tuned on a combination of QALB-2014 and ZAEBUC.

3.2.4 AES Module

The AES module leverages a fine-tuned version of CAMeLBERT-MSA to predict the CEFR levels of MSA essays. We fine-tune CAMeLBERT-MSA was on the ZAEBUC dataset and a larger synthetic dataset with topic, genre and level diversity for essay scoring (Qwaider et al., 2025).

[5] https://www.imarabe.org/
[6] https://www.arabacademy.com/alpt/

3.2.5 User Progress Tracking System

The User Progress Tracking System provides writers with clear and measurable feedback, recording CEFR scores and tracking error reduction over time. This historical data is presented through a linear graph that dynamically illustrates the user's learning trajectory.

3.2.6 User Profiling

Users who register have the option to input their native language or Arabic dialect and estimated proficiency level. This metadata enables more targeted prompting and feedback. It also supports further annotation of the auto-annotated essays collected to create, for example, (non-)native, dialect, or CEFR level specific profiles of users.

4 Preliminary User Experiment

Our goal is to determine if ARWI's feedback leads to measurable improvements in text quality such as reduction in grammatical errors or increased CEFR scores, and whether users find the UI/UX intuitive.

4.1 Experimental Setup

A total of 34 non-native mixed-gender undergraduate Arabic learners organized into four groups participated with proficiency levels ranging from A1-B1. Five essay prompts were offered tailored to the participants' CEFR level. Topics included Family and Friends, Sports and Hobbies, Spring Break, Travel Experience, and Weekly Schedule, with each essay suggested to be 120-500 words. A user survey was designed for UI and UX assessment, using a 5-point Likert scale with one-choice answers, along with two open-ended questions regarding the most and least useful features. Participants had 20 minutes for writing, 10 for corrections, and 10 for a user survey. A1 participants prepared texts in advance, allowing more time for correction.

Out of 112 total submissions, where users clicked the Check button and received feedback, 67 submissions were selected, representing the work of 12 different users, because they provided multiple submissions to incremental improvements to a single essay. 8 of these users reduced errors in their essay. One user submission contained only 3 errors in a 212-word initial draft but 4 errors in the final version, but with high CEFR scores suggesting this participant focused on content rather than error correction. The remaining submissions were by A1-B1 learners, where submissions typically contained tens of grammar errors.

Criteria	Avg. Score	Std. Dev
Clear navigation	3.68	0.90
User-friendly	3.71	0.89
Intuitive	3.59	1.09
Visually Appealing	3.03	1.03
Overall Satisfaction	3.65	0.58

Table 1: User feedback survey ratings regarding the UI experience. Ratings are on a 5-point Likert scale, with 5 being strongly positive, 3 neutral, and 1 strongly negative.

No instances of overall CEFR score improvement were observed during the 30-minute writing sessions. Significant score improvements on this relatively course-grained scale would likely require a much longer learning period.

The survey results shown in Table 1 indicate that the overall user experience of the system is moderately positive (see Appendix B for more details). Criteria such as "*Clear navigation*", "*User-friendly*", and "*Overall Satisfaction*" all received average scores around 3.65 to 3.71, suggesting that users generally find ARWI easy to navigate and use. However, the "*Visually Appealing*" criterion received a lower average score of 3.03, indicating room for improvement in visual design. Standard deviations (0.58 to 1.09) show a moderate degree of variability in users' perceptions, with the "Intuitive" rating exhibiting slightly higher deviation. This suggests that while many users appreciate the UI's intuitiveness, there is a subset for whom it is less clear. When asked whether they would recommend the system to others, approximately 85% of users responded affirmatively.

5 Conclusions and Outlook

By integrating a collection of essay prompts, a text editor, grammar error detection, correction suggestions, and automated essay scoring modules, ARWI provides targeted, iterative, actionable feedback that allows users to improve their writing and see improvements in their writing quality over time. We make ARWI publicly available at: https://arwi.mbzuai.ac.ae/.

Our preliminary experiment suggests the system is useful, but improvements are needed to the UI, a more fine-grained representation of progress would be useful, and more intuitive error correction hints are needed. We intend to incrementally improve the system based on further user experimentation, feedback, and analytics.

Limitations

Several aspects of ARWI require further refinement. The user interface needs adjustments based on user study feedback, including font size and screen real estate optimization. Error detection, classification, and correction suggestions require improved accuracy. Additionally, a larger study with a more diverse pool of native and non-native students across age groups, along with teacher feedback, is essential for a more comprehensive evaluation.

Ethical Considerations

The study parameters were approved by the internal review board (IRB) of our university. All user study participants were volunteers, and the purpose of the study was explained to them directly.

We recognize that AI assessment systems can make errors that may impact the student learning process and could be misused. This is not our intention. ARWI is designed to serve as a support tool for teachers and learners, not as a standalone evaluator.

References

A. Alfaifi, E. Atwell, and G. Abuhakema. 2013. Error annotation of the Arabic learner corpus. In *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg.

Abdullah Alfaifi and Eric Atwell. 2012. Arabic learner corpora (alc): A taxonomy of coding errors. In *The 8th International Computing Conference in Arabic*.

Bashar Alhafni and Nizar Habash. 2025. Enhancing text editing for grammatical error correction: Arabic as a case study. *Preprint*, arXiv:2503.00985.

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

C. o. E. Council of Europe. 2001. Common european framework of reference for languages: learning, teaching, assessment.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Tira Nur Fitria. 2023. Artificial intelligence (ai) technology in openai chatgpt application: A review of chatgpt in writing English essay. In *ELT Forum: Journal of English Language Teaching*, volume 12, pages 44–58.

Sylviane Granger. 1998. The computer learner corpus: A versatile new source of data for sla research. In Sylviane Granger, editor, *Learner English on Computer*, pages 3–18. Addison Wesley Longman, London & New York.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Minsun Kim, SeonGyeom Kim, Suyoun Lee, Yoosang Yoon, Junho Myung, Haneul Yoo, Hyunseung Lim, Jieun Han, Yoonsu Kim, So-Yeon Ahn, Juho Kim, Alice Oh, Hwajung Hong, and Tak Yeon Lee. 2024. Designing prompt analytics dashboards to analyze student-chatgpt interactions in efl writing. *Preprint*, arXiv:2405.19691.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar. Association for Computational Linguistics.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English. *Research Outputs*.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. *Preprint*, arXiv:2404.14914.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025. Enhancing arabic automated essay scoring with synthetic data and error injection. *Preprint*, arXiv:2503.17739.

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China. Association for Computational Linguistics.

Karin Ryding and David Wilmsen, editors. 2021. *The Cambridge Handbook of Arabic Linguistics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Lucía Sanz-Valdivieso. 2024. Technology-powered multilingual professional and technical writing: An integrative literature review of landmark and the latest writing assistance tools. *IEEE Transactions on Professional Communication*, 67(3):301–315.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, and et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, and et al. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

United Nations. 2024. Arabic language and AI: Advancing innovation while preserving cultural heritage. Accessed: 2025-02-28.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *Preprint*, arXiv:2303.13648.

K. Yancey, Geoffrey T. LaFlair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

A ARWI Interface

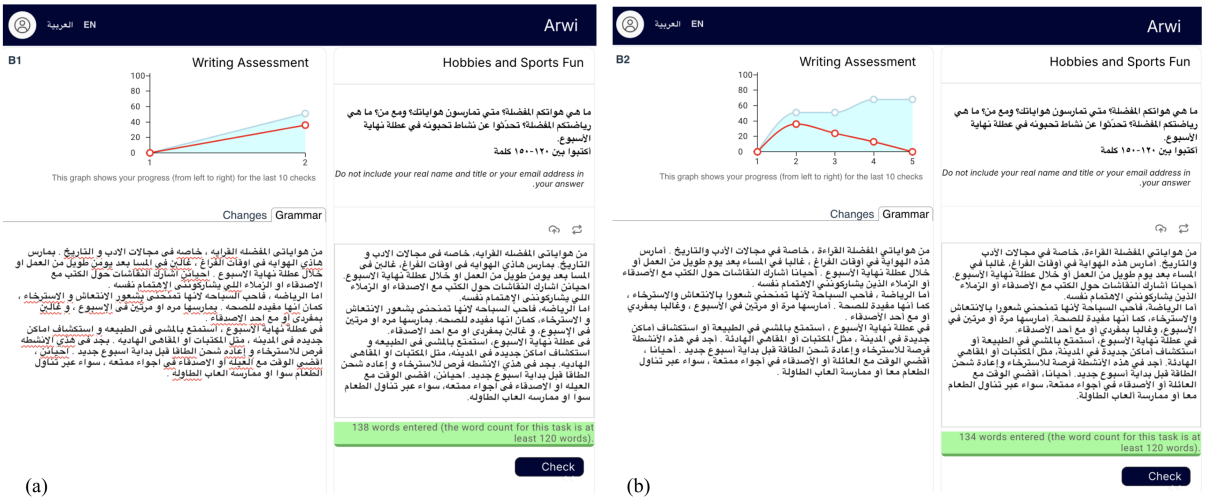

Figure 2: A before-and-after example of using ARWI's English interface. In (a) the text receives a B1 CEFR and a large number of errors marked with red underlining; in (b) the results shows improved writing and is raised to B2 CEFR. The **essay prompt** can be translated as "*What are your favorite hobbies? When do you practice your hobbies? And with whom? What is your favorite sport? Talk about an activity you enjoy on the weekend. Write between 120-150 words.*" The **written essay** can be translated as: "*One of my favorite hobbies is reading, especially in the fields of literature and history. I engage in this hobby during my free time, often in the evening after a long day of work or during the weekend. Sometimes, I participate in book discussions with friends or colleagues who share the same interest. As for sports, I enjoy swimming because it gives me a sense of refreshment and relaxation, and it is also beneficial for my health. I practice it once or twice a week, often alone or with a friend. During the weekend, I enjoy walking in nature or exploring new places in the city, such as libraries or quiet cafés. I find these activities to be an opportunity to relax and recharge before the start of a new week. Sometimes, I spend time with family or friends in a fun atmosphere, whether by sharing a meal together or playing board games.*"

B User Feedback Survey

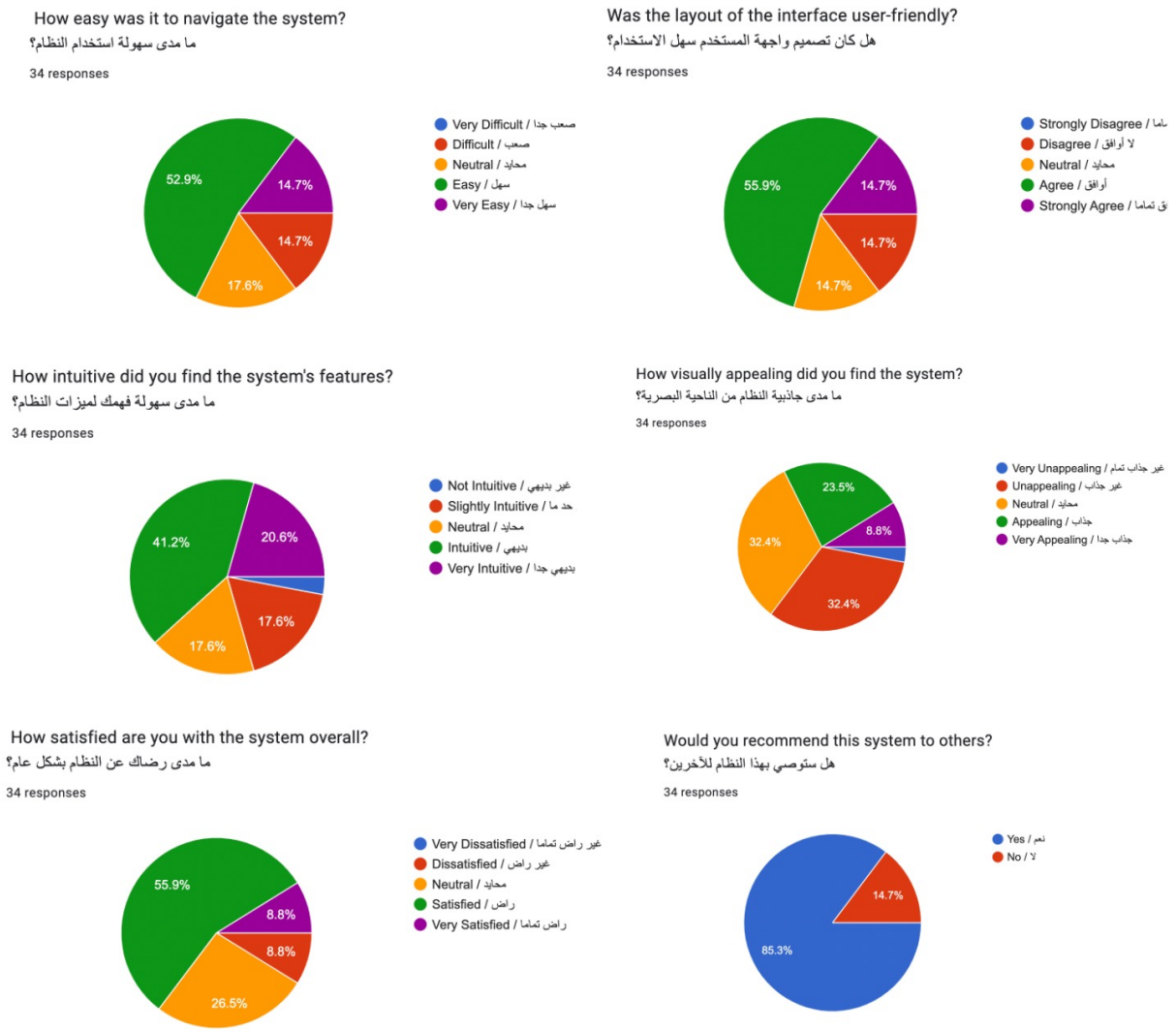

Figure 3: Qualitative feedback collected from 34 users who participated in the preliminary experiments with Arwi. The survey comprised five one-choice questions rated on a 5-point Likert scale and one binary question. The results highlight that certain aspects of the user interface–specifically its intuitiveness and visual appeal–require further refinement. Overall, users provided moderately positive feedback regarding their experience of usage.