ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Multi-Account Statistical Evaluation of ChatGPT Proficiency in the Kurdish Sorani Language

Alla Ahmad Hassan[1], Hemin Sardar Abdulla[2], Tara Yousif Mawlood[3], Rebwar Khalid Muhammed[4], Aso M. Aladdin[5]*, Tarik A. Rashid[6]

[1]Department of Database, Computer Science Institute, Sulaimani Polytechnic University, Sulaimani, Iraq, [2]Department of Software Engineering, College of Engineering and Computational Science, Charmo University, Sulaymaniyah, Chamchamal, Iraq, [3]Department of IT, Computer Science Institute, Sulaimani Polytechnic University, Sulaymaniyah, Iraq, [4]Department of Network, Computer Science Institute, Sulaimani Polytechnic University, Sulaimani, Iraq, [5]Computer Engineering Department, Tishk International University, Kurdistan Region, Iraq, [6]Computer Science and Engineering Department, University of Kurdistan Hewler, Erbil, Iraq

ABSTRACT

This research analyzes the strengths and weaknesses of ChatGPT in responding to questions posed in the Kurdish language, specifically its Sorani dialect, by evaluating its responses to a structured dataset of 50 multiple-choice questions across multiple topics such as language, history, culture, and general knowledge. Using four independent user accounts, each subjected to ten repeated testing cycles, the research assesses accuracy, consistency, and variation influenced by account identity, session timing, and model behavior. This study evaluates the multilingual capabilities of ChatGPT by comparing its performance in Kurdish (Sorani) and Arabic languages. The research establishes a framework to examine how artificial intelligence chatbots, such as ChatGPT, function as applied tools for language understanding and educational use. The analysis demonstrates that ChatGPT achieved an overall average accuracy rate of approximately 70%, indicating satisfactory performance in multilingual contexts. However, significant variations were observed across different user accounts, suggesting that factors such as user profile and temporal dynamics can considerably influence output consistency. The comparative findings highlight the developmental challenges in Arabic and Kurdish language processing, emphasizing the need for further refinement of ChatGPT's linguistic performance and its effective integration into academic and technological applications. While ChatGPT exhibited proficiency in answering general knowledge questions, it demonstrated a limited understanding of specialized topics in Kurdish, particularly classical literature and historical content. The research presents the strengths and limitations of ChatGPT for under-resourced languages and provides feedback to developers, educators, and researchers. Observing patterns in accuracy, question difficulty, and error behavior, this research also contributes to ongoing efforts toward improving the linguistic and cultural adequacy of AI models for under-resourced languages.

Index Terms: ChatGPT, Chatbot, AI, Kurdish Language, Sorani Dialect, Statistical Analysis

Access this article online

DOI: 10.21928/uhdjst.v9n2y2025.pp319-334

E-ISSN: 2521-4217
P-ISSN: 2521-4209

1. INTRODUCTION

Artificial intelligence (AI) is being developed across all sectors of life, and it is essential for all individuals to understand its significance, as it plays a crucial role in facilitating and streamlining various tasks [1]. As a result, the most advanced

Corresponding author's e-mail: Aso M. Aladdin, Department of Computer Science, College of Science, Charmo University, Sulaymaniyah, Chamchamal, Iraq. E-mail: aso.aladdin@chu.edu.iq

Received: 12-08-2025 Accepted: 08-11-2025 Published: 30-11-2025

definition describes AI as a collection of technologies [2]. These technologies enable computers to do a very abstract set of high-level things. These operations involve vision perception, language understanding and interpretation, data examination, decision-making, and others [3], [4]. One of the numerous uses of AI is creating software that will determine the best solutions and answers. These software programs assist in performing work and decision-making operations.

According to this, generally, a chatbot is a widely recognized software tool designed to replicate human-like discussions with users, typically using text or voice interactions [5]. Traditional chatbots rely on predefined rules and scripted responses, whereas more modern systems increasingly influence AI, particularly conversational AI approaches such as natural language processing (NLP), to read and respond to user input more effectively. These AI-powered chatbots can interpret user input in a more flexible and context-aware manner, enabling them to generate more accurate and dynamic responses, thereby improving user engagement and automating customer support more effectively [6]. Good chatbots can be incorporated into the organization's existing software to augment its applications. The addition of a chatbot to Microsoft Teams, for example, can make for a lively zone for content, tools, and team members to convene for chats, meetings, and collaboration [7]. Therefore, a major distinction and widely used classification between chatbots is based on the way they support users in conversations and produce responses. The former are rule-based chatbots, which usually present the users with a list of options to choose from [8]. These chatbots are often deployed in simple applications, like for frequently asked questions (FAQs). The second type is AI-based chatbots that are based on an AI, NLP, and machine learning (ML) [9].

AI chatbots are software tools designed to operate using various languages. These tools can process and respond to questions in multiple languages, though they often face limitations in supporting all linguistic variations and dialects, as well as different knowledge about different languages. Several AI programs, such as ChatGPT (by OpenAI), DeepSeek, Gemini, Microsoft Copilot, and Claude, have been developed to perform diverse functions for different purposes. In general, AI refers to computer systems capable of executing tasks traditionally associated with human intelligence in different languages, including learning, reasoning, problem-solving, perception, and decision-making [10]. As a core branch of computer science, AI focuses on developing algorithms that enable machines to interpret their environment, learn from data, and make intelligent decisions to achieve specific goals [11,12].

These AI tools, which rely on chatbots, can understand multiple languages; however, their accuracy varies depending on the language and the level of specificity required for the knowledge involved [13]. Each language presents unique challenges, ranging from grammatical rules and literary conventions to scientific terminology and cultural context. This includes knowledge about notable figures such as scientists, poets, novelists, and historical contributors, as well as the influence of various countries on the language and its development [14]. For the purpose of this study, the research focused on analyzing the fundamental differences in ChatGPT's responses to the same multiple-choice questions when presented in different languages, using separate accounts and at different times. The objective was to evaluate the accuracy and consistency of ChatGPT's answers across these languages and to identify any variations related to linguistic or dialectal differences.

Due to this complexity, achieving complete accuracy in any language is impossible. Therefore, each AI tool must be individually tested and evaluated for its error rate and performance in different languages [15]. Given this problem, our study examines the truthfulness of ChatGPT in answering questions across different domains of expertise in the Kurdish language chatbot with specific emphasis on the Sorani dialect. The motivation for this research stems from the limited quality of information available in the Kurdish Sorani language when using chatbot tools, particularly ChatGPT. During testing, this issue became clear. Moreover, no prior studies have addressed this problem or developed benchmarks for comparison. Therefore, this study proposes a framework to evaluate performance differences across languages using the same methodological approach. Although ChatGPT supports numerous languages, this research specifically focuses on comparing the Kurdish Sorani language with Arabic dialects due to their linguistic proximity. As a result, the aim was to assess the quality and accuracy of responses generated by ChatGPT. The primary objective of this research is to evaluate the degree of improvement in the accuracy of this AI tool, particularly in terms of how effectively it can address both general and language-specific queries.

The satisfaction users experience with an AI tool is strongly influenced by how warm, competent, and socially present the tool appears across different languages. Furthermore, the impact varies depending on the specific dialect, as using dialect-related information makes the chatbot more relevant and useful [16], [17]. In line with the rapid advancement of AI technologies, this study selected a variety of questions

across multiple domains of general knowledge, along with specific questions directly related to the Kurdish language.

In addition, the research also examines how accuracy percentages vary across different conditions. These factors include the user account, geographical location, and query time. It also considers the dynamic learning behavior of the AI when answering the same questions in different languages. As the AI framework stays learning and improving with time, such factors can determine the form in which answers are presented. Based on this context, the research contributes to a better understanding of ChatGPT's performance, providing information about its weaknesses and strengths in responding to questions in Kurdish language with Sorani dialect and wider knowledge domains. To achieve this goal, the study presents several key contributions, which are summarized in the following findings:

- Identified several chatbot categories and examined how dialects are used within multilingual processing. Chatbot systems, including ChatGPT, were tested to evaluate their performance. ChatGPT was used as a well-known AI tool to assess its limitations in answering questions in the Kurdish Sorani dialect. Its responses were compared with those in Arabic for the same queries. The findings showed that ChatGPT produced varied answers across different user accounts and time zones, indicating inconsistency in its output. Thus, the error ratio (ER) ranged from 39.4% to 24.8%, with the lowest value indicating different accuracy and reliable performance.
- ChatGPT showed gaps in knowledge for certain aspects of Kurdish and Arabic, particularly in historical topics, classical poetry, and prominent Sorani poets. To evaluate this, several multiple-choice questions were tested, and some had no correct answers across all accounts, indicating a consistent lack of knowledge in these areas. Statistical analysis examined ChatGPT's performance for each question, both in aggregate across all accounts and individually per account.

The study is structured as follows: Section 2 reviews multilingual language background, Section 3 outlines the methodology, Section 4 presents results and analysis, Section 5 discusses the findings, and Section 6 concludes with key insights and recommendations.

2. BACKGROUND REVIEW

2.1. Performance Evaluation of AI Tools

Although few-shot English learning has been advanced by in-context learning methods that do not update parameters [18], [19], few-shot cross-lingual transfer remains uncharted [20]. While encouraging results in non-English languages are seen [21], [22], [23]. However, it remains unclear whether these in-context learning methods can outperform traditional fine-tuning approaches across different languages and task types. This reflects a chronic lack of balance in NLP research, with English-biased development dominating and limiting development in low-resource and multilingual settings.

The BUFFET benchmark (Benchmark of Unified Format Few-shot Transfer Evaluation) attempts to mitigate this by evaluating few-shot transfer on 15 tasks across 54 languages [24]. Nevertheless, existing challenges such as data paucity, linguistic diversity, and test bias [25] [26] limit unbiased evaluation. Thus, advancement on multilingual benchmarks like ChatGPT still remains crucial to achieve unbiased performance beyond English.

In NLP, chatbots focus on parsing user input and generating responses, particularly by controlling prior knowledge. Several bots analyze the keywords and phrases used by users and gradually learn to respond more accurately and relevantly over time. In a chatbot system, it is not necessary for users to speak using the same dialect or linguistic structure as the chatbot itself [27]. When a user submits a message, the chatbot must interpret and handle the input [28]. However, since the user's language is often natural and unstructured, the chatbot requires a means to translate this input into a structured format that a machine can comprehend and act upon. As a result, this procedure is known as natural language understanding, as detailed under the dialect language component in Table 1. The conversion of a user's communication into structured data is referred to as constructing a semantic frame. This entire operation requires numerous steps, which are depicted in a general workflow in Fig. 1.

Furthermore, hybrid Chatbot [29] is a capable conversational agent which combines the advantages of rule-based systems and AI-driven approaches to develop a stronger and more versatile user experience. It combines accuracy and consistency achieved through the application of predefined rules for dealing with structured queries, with the understanding and responding capabilities to complex unstructured queries made possible by ML and NLP [30]. The combination of the two provides the chatbot with the ability to manage both types of interactions, from static FAQs to a more engaging, responsive, and contextually aware dialog, all while being more efficient, scalable, and providing better user satisfaction.

TABLE 1: Essential elements in chatbot design for multilingual and dialect-specific communication

Component	Purpose	Key applications	
NLU	A specialized branch of NLP that focuses on enabling machines to interpret, comprehend and derive meaning from human language. Unlike basic NLP, which deals with syntax and structure, NLU delves into intent recognition, contextual understanding, and semantic analysis, allowing systems to grasp nuances even across dialects.	AI chatbots and virtual assistants	Understanding user queries.
		Sentiment analysis	Detecting emotions in text.
		Intent recognition	Mapping user input to actionable tasks.
Dialect training data	Training a chatbot involves exposing it to vast amounts of relevant data and leveraging ML algorithms to help it understand, interpret, and respond to user inputs effectively.	Customer support	Handling FAQs, troubleshooting, and ticket routing.
		Virtual assistants	Managing tasks such as scheduling, reminders, and smart home control.
		Electronic commerce	Assisting with product recommendations and order tracking.
		Healthcare	Providing symptom checks and appointment bookings.
		Sentiment analysis	Detecting user emotions for better engagement.
Optional translation layer	Converts dialectal input into a standardized language form, enabling the chatbot to accurately interpret and respond to user queries.	Global customer support	Resolves queries in multiple languages without manual intervention.
		Multinational enterprises	Facilitates internal communications across global teams.
Response generator	Replies in standard or dialectal form, based on user need.	Intent recognition	Maps queries to actions (Dialogflow, Rasa).
		Contextual memory	Retains conversation history
		APIs/Webhooks	Fetches real-time data (inventory, weather) for accurate replies

AI: Artificial intelligence, NLU: Natural language understanding, NLP: Natural language processing, APIs: Application Programming Interface

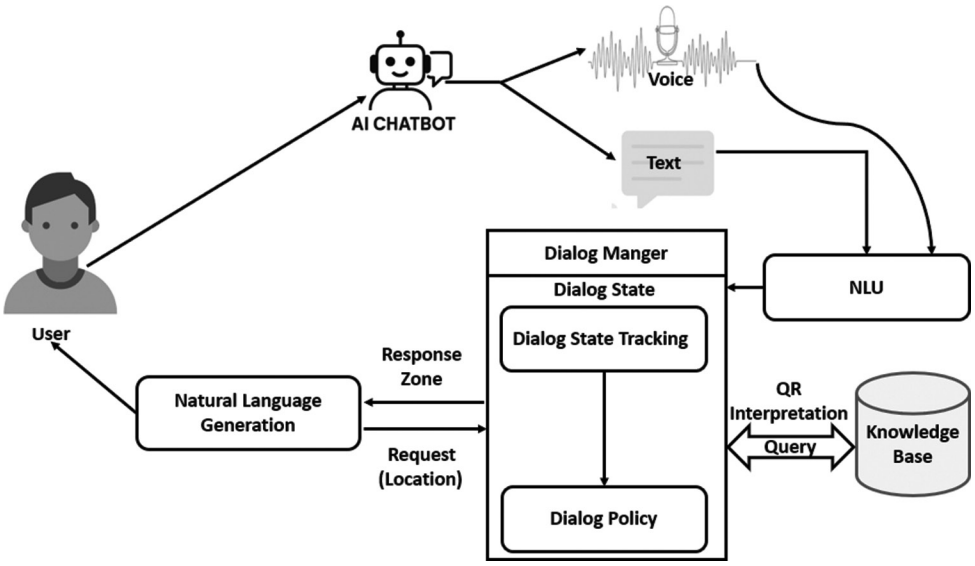

Fig. 1. The key elements involved in chatbot conversation processing.

To develop a multilingual chatbot with dialectal support, the system must include automatic language-detection modules, integrate translation APIs (Application Programming Interface), and be trained on datasets that contain diverse linguistic and dialect-specific samples [31]. Regional idioms and dialectal variations should be integrated through NLP

methods that enable appropriate comprehension and response within the parameters of the idiom or dialect from which they are expressed. To create or at least simulate a chatbot that understands dialects, it must possess the important elements given in Table 1.

As a result, the advantages and disadvantages of chatbot tools across different types have been recently surveyed. These studies highlight their potential to enhance personalized learning and student engagement. However, they also raise concerns, particularly regarding accuracy, cultural relevance, and ethical considerations.

2.2. ChatGPT Applications and Critical Analysis in Literature

The use of ChatGPT, an AI language model created by OpenAI, in Kurdish language education is a double-edged sword. Although ChatGPT performs well in languages with larger online corpora, its performance in less-represented languages, such as Kurdish, has not been as effective. It can serve as an autodidactic device, a dialog partner, and a creator of stimulating learning materials. This said, one should also be mindful of the challenges posed by engaging Chat GPT around human interaction, culture, mistakes and fixes, personal feedback [32]. Leveraging ChatGPT as a tool for learning the Arabic language in post-secondary education would provide fruitful opportunities in terms of learner experience on several platforms and within the existing operating systems and in terms of efficiency overall [33], [34].

It shows how, despite having the potential to improve language teaching by providing personalized feedback and tailored lessons, challenges posed by misinformation, absence of human interaction, and cultural sensitivity need to be solved, especially for less represented languages such as Kurdish [35]. Plus, in addition to the advantages of personalized learning and reduced tasks in education, ChatGPT is also a challenge due to the possibilities of academic dishonesty and issues related to identifying AI-generated text, thus threatening students' abilities to solve problems [36].

ChatGPT is an application of the transformer model for dialog generation and other NLP applications, such as question answering, developed by OpenAI. This is a language model for NLP tasks such as language generation developed by OpenAI [37]. It is capable of producing human-like conversational text. It is automatically generated, without human conversation or participation, chat in natural language, but it does not have human levels of understanding, empathy, or creativity [38]. Hence, ChatGPT is an OpenAI's state-of-the-art natural language model for natural conversational applications [39]. These applications and models, such as customer service, virtual assistants, and comparisons, use deep learning and NLP to provide human-like responses [39], [40]. ChatGPT, a generative AI chatbot, is another tool and resource that can aid language teaching and learning but must be used ethically and effectively with ethical and effective digital skills [41].

While ChatGPT may serve as a promising writing tutor for second language writing, more research is needed on the implications for writing pedagogy and for academic integrity [42]. Conversely, ChatGPT is a humanlike conversational agent based on the GPT architecture. Despite its limitations and ethical issues, it assists scientific studies by providing reviews of the literature, summaries of the content, and decisions in areas such as medicine and surgery [43].

The study examines the artificial language model ChatGPT, which generates human-like discursive text using deep-learning processes, and was developed by OpenAI. It engages as well with the opportunities and dangers associated with its use in an educational context [44]. The other looks into the use of ChatGPT in the item editing and modification process of high-stake testing, highlighting its use as an automated solution to a process within education and testing that essentially leads to enormous efficiency and cost benefits [45]. The model under investigation is ChatGPT, an OpenAI language model used for creative writing in poetry and prose. It explores how it works, what it takes into account when generating text, and the challenges of using it as a tool for creativity [46]. ChatGPT is a large language model for conversation developed by OpenAI. It is also human-like and biased because it has been trained on data generated by humans and then fine-tuned based on human feedback, which leads it to generate responses that are very much in line with what humans prefer and human cognitive biases [47]. The study analyses the application of GPT language models, namely ChatGPT, in higher education, focusing on their application as supports for engaging in theoretical exercises and laboratory practices and the questions of use and assessment in education. Effectiveness declines from theory to exercises to labs, as GPT struggles with specialized tools and software [48].

Critically, these literature studies present a balanced yet fragmented understanding of ChatGPT's role in education. Collectively, the studies highlight both its pedagogical benefits and ethical challenges. They emphasize potential for personalized learning, efficiency, and engagement but also reveal persistent limitations in accuracy, cultural sensitivity,

and academic integrity. Overall, the evaluations remain exploratory, requiring deeper empirical validation across diverse linguistic contexts.

3. METHODS

The present study takes a more systematic approach to gauge the accuracy and consistency of ChatGPT's responses to questions posed in the Kurdish language and specifically to the Sorani dialect. This process consisted of creating a set of multiple-choice questions, populating multiple test accounts with them, and exploring statistics on user performance. This approach also aims for cultural and linguistic relevance, testing reliability, and comprehensive evaluations along a variety of dimensions.

3.1. Data Collection

The study's dataset consists of 50 unique multiple-choice questions taken from a number of Kurdish language books. These questions are about geography, history, language structure, culture, and politics. There are four possible answers to each question, one of which is the actual answer. The data were handpicked to be culturally and linguistically relevant with the purpose of being useful for language assessment and educational use. The questions were hand-picked and screened for clarity, cultural, and linguistic appropriateness by native speakers of Kurdish in the Sorani dialect.

3.2. Data Preparation

Questions were assessed for clarity as well as for consistency in formatting. Response options were identified using the Kurdish alphabet (*a,b,c,d*), and all text was rendered in Unicode to help the display of the Kurdish script properly. The questions were entered into a database in a structured format suitable for processing and analysis.

The present study employs already established assessment frameworks through four distinct test accounts to measure ChatGPT's performance in the Sorani dialect Kurdish language. Each account contains 50 diverse questions that test different aspects of the knowledge of the Kurdish language. To obtain reliable, robust data, each question was asked ten times. The procedures for question formulation and the criteria employed for evaluating the responses are comprehensively outlined in the following section.

Fig. 2 shows the workflow of the experimental design to evaluate ChatGPT's performance in answering Kurdish (Sorani) questions. It starts with four independent ChatGPT accounts (A, B, C, D), which all receive the same set of 50 questions. Each account goes through the processing of the questions over several testing sessions. All the responses are kept in separate data sets and statistical analyses are performed on it. The result is an analysis of ChatGPT's correctness, types of errors, and consistency of performance, which offers new observations regarding how knowledge is represented by the model in the context of a low-resource language.

3.3. Experimental Design

To analyze the consistency and reliability of ChatGPT, four separate ChatGPT accounts (A, B, C, and D) were used in

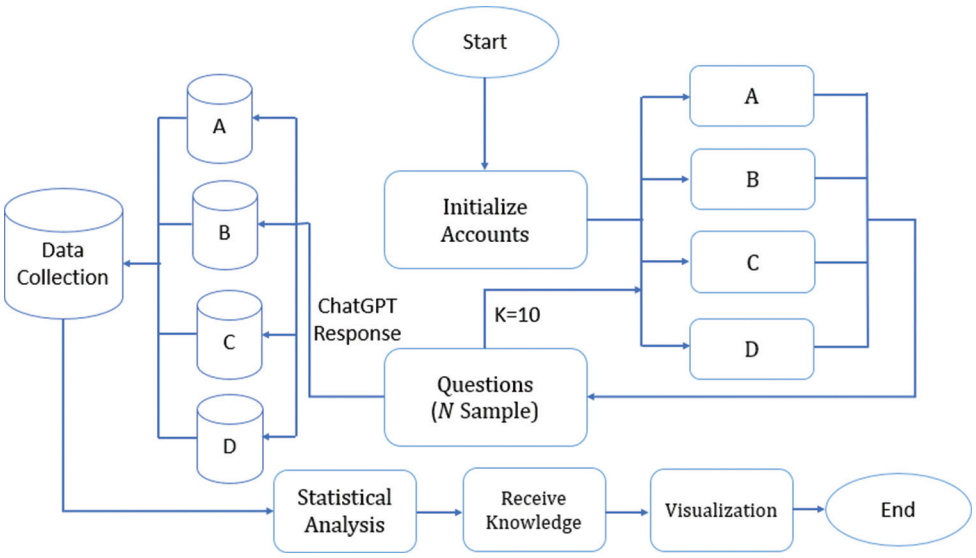

Fig. 2. Multi-account testing and data collection workflow.

the study. Both accounts underwent ten test runs on the complete dataset of 50 questions. This resulted in a total of 2000 responses (4 accounts × 10 sessions [Rounds] × 50 questions). Testing for each iteration occurred at varying times of day for each account over a total period of five days. Tests were either scheduled for morning hours (between 6 AM and 9 AM GMT) or evening hours (between 6 PM and 12 AM GMT). Questions were presented in a randomized order for each session to prevent any learning or memorization effects. The answers were all transcribed in a file for each account and session, which were ChatGPT's responses in each dialect of the Kurdish language. The questions, as noted above, were scientific, literary, and geographical. The process is depicted in Fig. 3. Then, the next step is to conduct the evaluation, which is described in the next sections, in terms of how the responses perform and what they look like.

3.3.1. Question-answer evaluation based on test rounds collections

As determined, the process begins with the initialization of four independent accounts (A, B, C, and D), each configured to perform the same set of tasks under identical conditions. The core of the research involves a loop that repeats ten times, ensuring consistency and reliability throughout the repeated testing process. Within each loop iteration, all four accounts simultaneously process a standardized set of 50 questions, designed to evaluate AI performance. The method reflects a systematic, repeatable testing structure that enables both cross-account comparison and longitudinal performance evaluation, contributing to the methodological rigor of the study.

According to this, to evaluate the responses generated by ChatGPT, a set of fifty questions was used, distributed across 40 test sessions conducted through four different user accounts. These sessions are used to assess the average performance of ChatGPT, with attention given to identifying any variations based on the time of testing or the specific user account. The evaluation results are presented using equation 1 as the average score per session or rounds, which highlights the performance outcomes for each individual test session.

$$Average = \frac{1}{n}\sum_{i=1}^{n} T_i \tag{1}$$

In this evaluation, n represents the total number of test sessions conducted, while T_i refers to the number of correct answers provided by ChatGPT during the *i–th* session. Each session corresponds to one instance of testing, and the value of T_i reproduces the model's accuracy within that specific session.

3.3.2. Question-answer evaluation based on question-level collections

Each question was answered in forty separate test sessions, resulting in a set of forty responses per question. This allowed for a comprehensive tracking of each question's performance across all test cycles. By analyzing these results, it was possible to identify which questions were consistently answered correctly and which were frequently misunderstood. This evaluation supported the classification of questions based on their difficulty and clarity, offering insights into how ChatGPT interprets and adapts its responses, especially when working with variations in question structure across different tests.

This method proved especially useful for assessing ChatGPT's accuracy in responding to questions in the Sorani

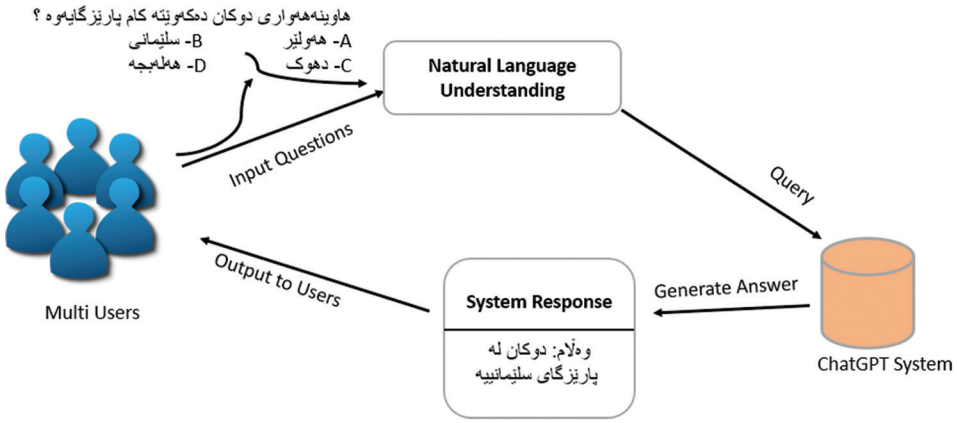

Fig. 3. The multiuser framework for question answering in Sorani dialect with ChatGPT.

dialect. Equation 2 was used to calculate the truth ratio (TR), representing the proportion of correct answers for each question or test session. Following this, Equation 3 was applied to determine the proportion of incorrect responses (ER), providing a balanced view of the model's strengths and limitations in handling Sorani language queries.

$$TR = \frac{T}{N} \tag{2}$$

In this context, T represents the number of correct responses, also referred to as "True" answers. N denotes the total number of responses collected. These two values are essential for calculating the accuracy or TR of the model's performance, as they allow for a straightforward comparison between the number of correct answers and the overall number of attempts.

$$ER = \frac{F}{N} = 1 - TR \tag{3}$$

In this evaluation, F represents the number of incorrect responses, also referred to as "False" answers. This value is used to measure the frequency of errors made by ChatGPT during the test rounds for each question. Question accuracy across all accounts is evaluated to determine how well each individual question performed throughout the test sessions. This is done using Equation 4, which calculates the accuracy of each question by dividing the number of correct responses by the total number of attempts. Specifically, T_j represents the number of correct responses to question j, and R_j denotes the total number of attempts for that same question. This measure provides insight into which questions were consistently understood and answered correctly by ChatGPT, helping to identify patterns in question difficulty and clarity.

$$Accuracy_{questionj} = \frac{T_j}{R_j} \tag{4}$$

3.3.3. Methodological validation framework

The multi-account, multi-round testing framework applied in this study was systematically structured and grounded in established scientific validation principles. Specifically, the design follows concepts of test–retest reliability and cross-account validation derived from psychometric evaluation and computational linguistics. These approaches are commonly employed to assess temporal consistency, reproducibility, and inter-sample stability in language evaluation research. Moreover, the quantitative assessment using TR, ER, and accuracy per session (Equations 1–4) ensures methodological rigor, transparency, and reproducibility by quantitatively verifying the stability and reliability of the model's responses across independent test accounts and repeated trials. This alignment establishes that the present evaluation framework adheres to recognized scientific standards of validity and reliability.

4. RESULTS AND ANALYSES

In response to the rapid advancement of AI technologies, this study systematically selected a comprehensive set of questions encompassing a wide range of general knowledge domains, alongside specialized items specifically pertaining to the Kurdish language. Furthermore, this study contributes to a more nuanced understanding of ChatGPT's performance, highlighting both its capabilities and its limitations in addressing inquiries formulated in the Sorani dialect of Kurdish, as well as in broader domains of general knowledge.

We fully understand the concern regarding the adequacy of using 50 multiple-choice questions to evaluate a large language model like ChatGPT. In our study, these 50 carefully selected questions were tested across four independent ChatGPT accounts, with each account completing ten repeated sessions under different conditions (locations, time zones, and testing rounds), resulting in 2,000 total responses ($4 \times 10 \times 50$). The repetition of the same questions across multiple accounts and sessions was a deliberate methodological choice aimed at examining consistency, reliability, and temporal variation in ChatGPT's responses, rather than solely its breadth of knowledge. Interestingly, the results revealed that even when the same question was repeated, the model did not always provide the same answer across accounts or test rounds. For example, questions related to Kurdish classical poetry were not answered correctly by any of the accounts, highlighting a consistent gap in ChatGPT's knowledge of culturally specific content. This design allowed us to identify not only performance differences between accounts but also systematic weaknesses in certain knowledge domains. We will consider expanding the number and diversity of questions in future studies to provide broader coverage while maintaining cross-account validation.

4.1. Test Rate Evaluation

The comparison study conducted across the four test accounts, each exposed to ten repetitions per question, demonstrates a statistically significant variation in the consistency of response performance. These findings indicate the necessity for focused methodological interventions and model

optimization tactics to enhance the overall dependability and accuracy of the system's outputs.

Test results show that ChatGPT was more correct than incorrect answers, showing that this tool can be relied upon. Table 2 presents each account's average number of true and false responses to 40 test questions, as well as each account's ratio of true to false responses. As for the average of the tests, it is higher than the average of the true answers, being 34.875 and 0.6975, respectively, which indicates a strong tendency toward the correct answers. The lower average of 15.125 with a false response ratio of 0.3025, on the other hand, indicates false answers.

To show the direction of the TR and the ER, 40 tests were performed on a set of four accounts: A, B, C, and D; each of which was tested ten times over the course of 5 days, as outlined in the methods section. The TR starts relatively high at about 70% but then significantly drops between the first and the early tests where it approximates 55% in test 4. The ER, on the other hand, starts at around 30% and reaches its maximum of approximately 50% around there. Beginning with test 7, there is a dramatic turn in that the TR is significantly higher and remains consistent between 70% and 80% with the ER dropping low and stabilizing between 20% and 30%. The analysis maintains this correlation between the two ratios. The ratios are seen to stabilize during the tests, showing that the performance level is reached. The higher TR and lower ER, achieved posttest 7 in Fig. 4, indicate that the subject was overcoming some difficulties.

TABLE 2: The average response to all test questions

Test case	Test average	Ratio
True answer	34.875	0.6975
False answer	15.125	0.3025

Four accounts were compared in terms of their reaction to a series of test questions, and the results differed both in accuracy and the rate of errors. The analysis is based on the average number of true and false answers given by each account, along with the TR and the ER percentages, respectively. There is also an overall average for all four accounts to provide a point of comparison. A segmentation for this breakdown is presented in Table 3.

The performance of four different accounts has been evaluated based on their responses to a series of test questions, revealing notable variations in accuracy and error rates. When examining the true answers, the average number of correct responses increases progressively across the accounts. *Account_A* demonstrates the lowest average with 30.3 correct answers, while *Account_D* achieves the highest with 37.6. This steady improvement indicates a gradual enhancement in performance from *Account_A* to *Account_D*. Conversely, the false answers show an inverse pattern: The average number of incorrect responses decreases from 19.7 for *Account_A* to 12.4 for *Account_D*. This inverse relationship between correct and incorrect answers supports the conclusion that *Account_D* performs the best, while *Account_A* leftovers the most.

Further evidence is provided by the TR and ER, which reinforce these observations. The TR, representing the percentage of correct answers, increases from 60.60% for *Account_A* to 75.20% for *Account_A*. Similarly, the ER decreases from 39.40% to 24.80% over the same range.

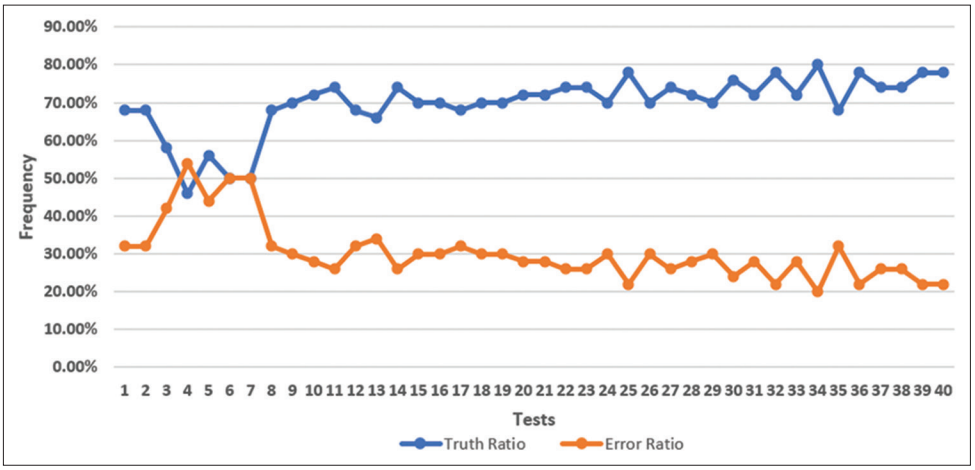

Fig. 4. The frequency of correct and incorrect responses across all tests.

TABLE 3: The average responses to test questions for each account

Test case	*Account_A*	*Account_B*	*Account_C*	*Account_D*	All
True answers	30.3	35.1	36.5	37.6	34.875
False answers	19.7	14.9	13.5	12.4	15.125
Truth ratio	60.60%	70.20%	73.00%	75.20%	0.6975
Error ratio	39.40%	29.80%	27.00%	24.80%	0.3025

These trends confirm the relative performance levels of each account and highlight the potential for targeted improvements. An overall average across all accounts is also included to serve as a baseline for comparison.

The performance of the four accounts has been tested by trial on ten consecutive average test questions and has been concerned with the ratio of correct to wrong answers. The comparison provides us with insight into how well each account performed over time and emphasizes fluctuation in accuracy. A clear visual representation of these trends is provided in Figs. 5 and 6. The primary purpose of these figures is to visually represent the fluctuations in the ratio of correct responses for each account, providing a comparative analysis of their performance over time. The proportion of correct responses is measured between 40.00% and 85.00%, and the testing axis indicates the time-wise arrangement of the ten tests. The performance of each account is plotted separately in a way that an explicit comparison and contrast could be shown for the test period.

Account_A exhibits the highest level of performance variability among the four accounts. It begins with an accuracy of approximately 70%, followed by a sharp decline to below 50% by the fourth test. This low level persists until the eighth test, after which a sudden rebound restores its performance to the initial range. Such fluctuations indicate a lack of stability in *Account_A*'s performance and may reflect underlying issues affecting its ability to generate accurate responses. In contrast, *Account_B* maintains a relatively stable accuracy throughout the testing period, fluctuating within a narrow range of 65% to 75%. Minor rises and dips are observed, but overall, the account demonstrates consistent performance with minimal variation, indicating reliability over time. In addition, *Account_C* also performs at a relatively high level, with accuracy predominantly exceeding 70%. Although it shows some moderate fluctuations, including a slight decline during the middle phase of testing, it recovers toward the end. This pattern suggests a generally resilient performance, with temporary setbacks that do not significantly impair overall accuracy. Finally, *Account_D* consistently achieves the highest performance among all

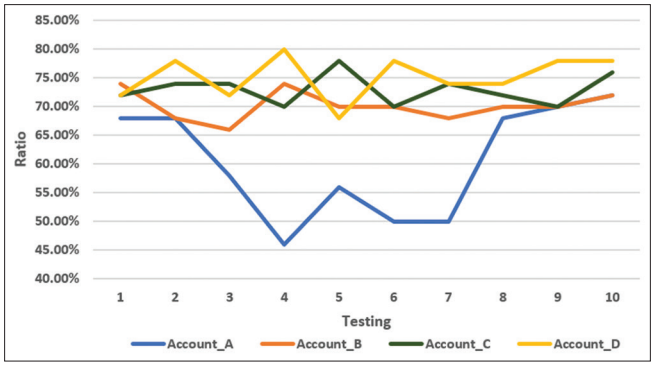

Fig. 5. The ratio of correct responses based on account testing.

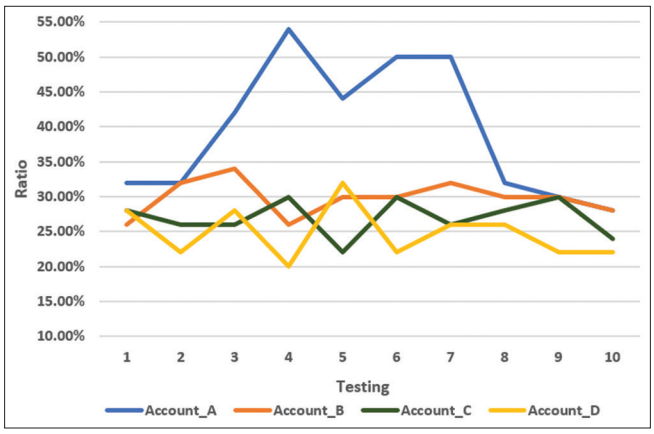

Fig. 6. The ratio of incorrect responses based on account testing.

accounts. Its accuracy remains above 70% across all tests, peaking at approximately 80% during the fourth test. While there are minor variations, its overall trend reflects a stable and robust level of accuracy, indicating dependable performance across the entire testing period.

These results highlight that each account produced different outcomes in response accuracy when evaluated using test questions in the Kurdish Sorani dialect through ChatGPT tools or different Chatbots. This fluctuation implies that AI performance is not uniform across accounts and may be influenced by training discrepancies or disparities in request zones. Consequently, these findings underline the need for

continued refining of AI technologies to ensure uniform and equal performance across languages and user scenarios.

4.2. Questions Rate Evaluation

Nonetheless, the use of interpreters for every dialect significantly impacts the chatbot responses and is an essential factor in training and handling user inputs of multiple languages. As a result, the system occasionally gives different responses to a repeated question of which some might be correct and the others not correct. Inconsistencies typically occur when the dialog manager tries to give responses without enough information about the context of the question.

In particular, ChatGPT exhibits notable limitations in handling the Kurdish language, particularly in the context of Kurdish Sorani literature, poetry, and the knowledge of Kurdish poets. This subsection evaluates the response patterns from four labeled accounts, focusing on how their answers vary, sometimes randomly, when presented with specific questions related to Kurdish Sorani. The discrepancies point toward the reality that ChatGPT does not have adequate knowledge in such domains and therefore generates faulty outputs. Therefore, this analysis provides valuable insights and proposes future research for enhancing AI performance on underrepresented languages. It also requests enrichment of training data sets and language resources, especially for a dialect like Kurdish Sorani, to render future AI responses more accurate and culturally appropriate.

The discrepancies point toward the reality that ChatGPT does not have adequate knowledge in such domains and therefore generates faulty outputs. Therefore, this analysis provides valuable insights and proposes future research for enhancing AI performance on underrepresented languages. It also requests enrichment of training data sets and language resources, especially for a dialect like Kurdish Sorani, to render future AI responses more accurate and culturally appropriate.

Accordingly, the valuation of results highlights how frequently various feature test cases occurred across all accounts, providing a quantitative view of scenario distribution and performance levels. Test case identifies each scenario of counting the number of question-response (QR), frequency shows its occurrence count of responses according to the testes, and ratio reflects its percentage of the total. All summarized data are shown in Table 4. In the feature test, all questions were answered correctly in 18 tests, representing 36% and incorrectly in 3 tests, representing 6%. These three questions, which consistently receive incorrect answers, are related to poems, poet names, and historical knowledge. This significant discrepancy suggests either a high level of accuracy in other areas or a strong bias toward providing correct responses elsewhere.

TABLE 4: The frequency ratio of feature test cases across all accounts for each question.

Test case	Frequency	Cumulative frequency	Ratio (%)	Cumulative percent (%)
No. of all QR truth 40	18	**18**	36	36
No. of all QR false 0	3	21	6	42
No. of QR up to 30 truth	13	34	26	68
No. of QR up to 20 truth	4	38	8	76
No. of QR up to 10 truth	6	44	12	88
No. of QR up to 0 truth	6	50	12	100

QR: Question-response

Moreover, the subsequent segmentation of test cases is based on the number of correct answers within a defined threshold. In particular, cases where the number of correct answers ranged from 30 to 39 occurred 13 times, accounting for 26% of all tests. Although this result based on all questions is considered acceptable, the outcome when considering only the questions with all query-related responses marked as true reaches approximately 61%. This indicates that achieving optimal performance does not necessarily lead to a higher chance of correctly answering the selected query-related questions.

Test cases with a maximum of 20 correct ones were seen 4 times, and these comprised 8% of all tests. Cases with a maximum of 6 and a maximum of 0 correct ones each also appeared 6 times, representing 12%, respectively. These categories enable us to take a more general look at distribution in performance, with differing rates of accuracy across accounts. Notably, earning up to 30 accurate answers showed the second-highest frequency after full-score situations, validating the observation of generally excellent accuracy. The results connected to QRs provide insight into performance trends and accuracy across tests, demonstrating the overall reliability of the tested accounts. However, many restrictions remain, particularly in generating accurate replies in Sorani Kurdish. This reflects limitations in ChatGPT's handling of some Kurdish dialects and domain-specific information, especially when striving for optimal answers across multiple user accounts and geographies.

One of the cases in the study illustrates the TR and ER for all questions across all accounts, based on the frequency of correct and incorrect QR responses over a set of 50 questions. The primary objective is to visually demonstrate the variation in the percentage of correct and incorrect answers across this range, as shown in Fig. 7.

It can be observed that with high TR, ER will be low, and low TR will result in high ER, which models an anticipated reverse relationship between correct and erroneous QR. The measures will normally spike to 100%, which indicates complete accuracy, yet fall to 0%, which indicates complete inaccuracy. Such extreme fluctuation reflects significant variation in performance, featuring normal swings between absolute accuracy and absolute error, consequently implying erratic response quality between accounts.

The analysis provides a comparison of feature test cases in four accounts. The main intention is to describe the distribution of each account's test result, including information about their relative performance and stability. It was found that QR frequencies may differ between accounts, with time zone and domain being potential deciding factors. Thus, the division is by the frequency at which each test case appeared by account or by the frequency at which each given result occurred. These results are summarized in Table 5.

The findings categorized the cases according to the total number of correct answers to QRs per account. *Account_B* has the highest count at 30 correct answers out of a total of 50 questions answered correctly, followed closely by *Account_C* at 28 and *Account_D* at 27 correct answers. *Account_A* has a relatively lower frequency of 18. For cases of all questions being answered wrongly, *Account_B* leads with 7 while *Account_A* indicates 6 and *Account_C* and *Account_D* each indicate 5. This implies that while *Account_B* is the one with the highest number of perfectly correct answers, it also indicates a relatively higher rate of overall failure among the accounts. In addition, Table 5 segment splits tests into various groups based on the number of QRs in given intervals. *Account_D* has at most 5 correct test scores with 9; *Account_A* and *Account_C* have 8 tied, and *Account_B* has 3. *Account_A* and *Account_C* have <5 correct test scores with 10 and 9, respectively, *Account_B* has 9, and *Account_D* has 5, meaning *Account_D* with fewer low scores. For more precise 5 correct answers for examinations, the top frequency is in *Account_A* with 8, then *Account_D* with 4, *Account_B* with 1, and *Account_C* with zero, indicating high heterogeneity across accounts.

The frequency of correct responses across four different accounts over a series of 50 questions is illustrated in Fig. 8. The graph provides a visual comparison of how the number of accurate answers varies for each account, highlighting differences in performance throughout the set. Conversely, Fig. 9 displays the opposite of Fig. 8, indicating the frequency of wrong replies across the QRs for the four different accounts. The frequency reflects the number of correct or

Fig. 7. The rate of correct and incorrect responses for all questions across all accounts.

TABLE 5: The frequency ratio of feature test cases across different accounts

Test cases	*Account_A*	*Account_B*	*Account_C*	*Account_D*
No. of all QR truth 10	18	30	28	27
No of all QR. false 0	6	7	5	5
No. of QR up to 5 truth	8	3	8	9
No of QR less than 5 truth	10	9	9	5
No of QR equal to 5	8	1	0	4

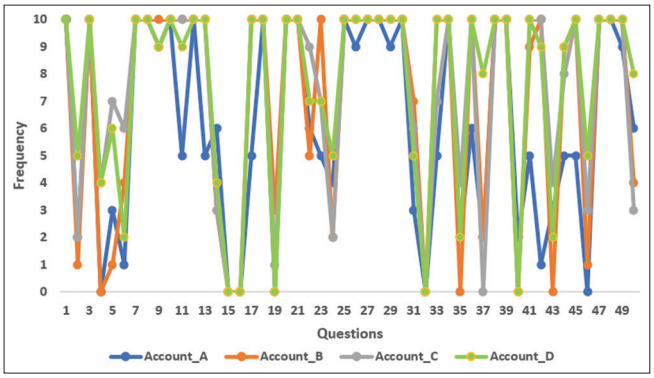

Fig. 8. The frequency of truth responses for each different account.

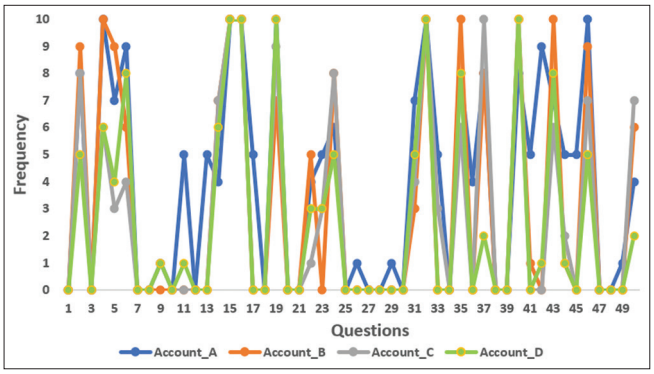

Fig. 9. The frequency of false responses for each different account.

incorrect responses, ranging from 0 to 10 tests per account over 50 distinct questions. Each account's performance is represented independently, allowing for easy separation and comparison throughout the question sequence.

The response patterns of the four accounts are very different in consistency and response correctness. *Account_A* is the most variable with large oscillations between 0 and 10 correct responses, which show extreme inconsistency although its correct responses show fewer errors than others. *Account_B* is also fluctuating but less wildly and in a pattern that indicates a slightly more stable but still fluctuating performance. *Account_C* has only moderate volatility, with most times having a higher proportion of correct responses with fewer instances of zero, indicating more stable accuracy. *Account_D* is characterized by the most stable and consistently high level of performance, staying very close to 10 correct responses on all accounts, indicating solid and consistent ability.

The results clearly show that the interpretation of queries by chatbots using ChatGPT varies significantly, particularly when different dialects are involved. This variation is especially evident in the test queries analyzed during this investigation. The accuracy and consistency of responses differed depending on the dialect used, highlighting challenges in understanding and processing language variations. In addition, external factors like the time zone that the test was aimed at and response time per question also contributed. These could have skewed the performance of the chatbot, leading to variations in results between accounts.

5. DISCUSSION

As demonstrated in the previous section, the analysis focused exclusively on evaluating responses in the Kurdish Sorani language. This area lacks prior research and serves as a new benchmark for assessing ChatGPT's performance compared with earlier versions. The identified limitations highlight the need for deeper analysis by comparing responses in Kurdish with those in Arabic using the same set of questions. Such a comparison provides a structured framework for linguistic evaluation. The findings also support the academic value of this framework in education, showing how AI tools can enhance teaching methods, classroom engagement, and lesson planning.

The comparative model of Kurdish and Arabic results has significant ramifications for the academic use of ChatGPT. Kurdish responses are both less truth-accurate, 69.75% and more error-prone, 30.25% than Arabic, 73.6% truth, 26.4% error on four accounts. This variation reflects language-resource disparities as well as model optimization differences for Kurdish as a low-resource language. Researchers can utilize these findings to understand ChatGPT's disparate performance across languages, with realistic classroom use expectations. For classroom use, Arabic users are able to anticipate more accurate content generation, while Kurdish instructors must utilize critical validation and further clarification to achieve content quality. The results indicate ChatGPT's promise as a complementary pedagogical tool for idea generation, translation, and text support but emphasize language-based testing and ethical revision to prevent misinformation and ensure inclusivity in linguistically diverse classrooms. These results are clearly reflected in the benchmark comparison presented in Table 6.

The statistical comparison between Arabic and Kurdish language answers in Table 7 recognizes that the two languages have similar patterns, although the differences are not conclusive but recognize major challenges for AI optimization when it works in multilingual settings. Kurdish

TABLE 6: A framework comparison between two languages (Kurdish and Arabic) during answer generation across four selected ChatGPT accounts

Test case	*Account_A*		*Account_B*		*Account_C*		*Account_D*		All	
	Kurish Lang	Arabic Lang	Kurish Lang	Arabic Lang	Kurish Lang	Arabic Lang	Kurish Lang	Arabic Lang	Kurish Lang	Arabic Lang
True answers	30.3	31.8	35.1	36.2	36.5	38.9	37.6	40.3	**34.875**	**36.8**
False answers	19.7	18.2	14.9	13.8	13.5	11.1	12.4	9.7	**15.125**	**13.2**
Truth ratio	60.60%	63.60%	70.20%	72.40%	73.00%	77.80%	75.20%	80.60%	**0.6975**	**0.736**
Error ratio	39.40%	36.40%	29.80%	27.60%	27.00%	22.20%	24.80%	19.40%	**0.3025**	**0.264**

TABLE 7: ChatGPT response accuracy and frequency comparison between Kurdish-Sorani and Arabic languages

Test case	Kurdish Lang			Arabic Lang.		
	Frequency	Ratio (%)	Cumulative percent (%)	Frequency	Ratio (%)	Cumulative percent (%)
No. of all QR truth 40	18	36	36	19	38	38
No. of all QR false 0	3	6	42	4	8	46
No. of QR up to 30 truth	13	26	68	15	30	76
No. of QR up to 20 truth	4	8	76	5	10	86
No. of QR up to 10 truth	6	12	88	3	6	92
No. of QR up to 0 truth	6	12	100	4	8	100

answers were negligibly lower in TRs in some cases, while Arabic answers underwent mild changes, which depicts ChatGPT's less consistent management of low-resource languages like Kurdish. These variations cause high-impact problems in learning use cases with content requirements for precision and reproducibility. To counteract this, future optimization of ChatGPT needs to focus on enhancing its understanding of Kurdish morphology, syntax, and context semantics with maintenance Arabic accuracy. Providing the AI with the same standardized questions across each language, alongside properly constructed multilingual data sets, can aid in tuning performance and reproducibility. In education, these multilingual AI materials can be used to prepare lessons, classroom lectures, and homework, offering equal learning chances for every language. Table 7 indicates comparative results, showing highly distinguishable differences in truth frequency and ratio that result in targeted improvements.

6. CONCLUSION

This work presents a comparative analysis of ChatGPT's performance in responding to multiple-choice questions (QR sets) in the Kurdish Sorani language within a methodologically controlled testing environment. The research also compares these responses with Arabic-language QR sets, rendering it feasible to critically test ChatGPT's performance across different multilingual settings. The technique provides valuable suggestions to users, particularly lecturers in higher learning institutions, on how best to utilize ChatGPT and also reveals potential limitations for developers to address. Through the experimentation of 50 expertly drafted questions on four user accounts over a number of sessions, the study reveals the strengths and weaknesses of ChatGPT as a tool for underrepresented languages. Some of the questions had no proper answers in all accounts, reflecting systematic deficiencies in the model's understanding. Interestingly, wrong answers were not randomly produced; instead, the AI reused the same mistakes again and again, exhibiting recognizable patterns in its performance instead of guessing.

The overall accuracy rate of around 70% reveals an effective baseline of performance, demonstrating that ChatGPT, when employed through chatbots in multiple interpretation situations, displays a core comprehension of the Sorani dialect and general knowledge themes. However, the results also highlight substantial differences across different accounts and test sessions, suggesting that ChatGPT's effectiveness can be influenced by factors such as time, session context, and user-specific activities. These findings emphasize the importance of evaluating AI systems not only in terms of average accuracy and statistical comparisons between languages, such as Kurdish and Arabic, but also with respect to their consistency and reliability across different testing conditions.

Furthermore, the study highlights significant limitations of the model in handling culturally specific or historically nuanced

content, especially in areas such as classical Kurdish literature and regional history. The gap itself is a general issue in AI design around the underrepresentation of minority cultures and languages in training datasets. While the model did not make completely random mistakes and would replicate very similar errors, it did not manage to be as profound in sense as it needed to be for more specialized fields.

Even within such constraints, a steady increase in trends of the same type of response rates in repeated testing holds promise for adaptive learning or higher convergence toward user input in the long term. The study provides a balanced framework of appraisal through its multi-layered analysis, consisting of session trends, account comparisons, and QR assessments, and provides an accurate measure to examine AI performance under low-resource language conditions. In conclusion, while ChatGPT shows promise as a tool for language learning and information retrieval in Kurdish or Arabic, its current limitations warrant cautious integration, especially in educational or culturally sensitive applications. Future work enhancements need to go toward incorporating linguistic datasets, enhanced cultural training, and context-aware reasoning.

The conclusions drawn from this study provide valuable insights not only for developers and educators working with Kurdish but also for the broader discourse on ethical and inclusive AI development for marginalized languages and cultures. While the chatbot operates with a limited dataset and within a security framework utilizing symmetric algorithms, several limitations remain, highlighting areas for further improvement. Despite these constraints, the findings demonstrate the potential of ChatGPT in educational contexts, supporting teaching and learning activities through structured QR interactions [49]. Furthermore, the study underscores its applicability in evaluating the effectiveness and impact of AI chatbots in other domains, such as healthcare, particularly when assessed through QR-based testing. Overall, these results reinforce the importance of adapting AI systems to low-resource languages while maintaining ethical, reliable, and culturally inclusive practices. In addition, the differences between ChatGPT-4.0 and the newly released ChatGPT-5 can be evaluated through a future comparative study.

REFERENCES

[1] N. Rane. "ChatGPT and similar generative artificial intelligence (AI) for smart industry: Role, challenges and opportunities for industry 4.0, industry 5.0 and society 5.0". *SSRN Electronic Journal*, vol. 2, no. 1, pp. 10–17, 2024.

[2] K. Ofosu-Ampong. "Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions". *Telematics and Informatics Reports*, vol. 14, p. 100127, 2024.

[3] M. J. Sousa, S. Pani, F. dal Mas and S. Sousa. "*Incorporating AI Technology in the Service Sector*. Apple Academic Press, New York. 2024.

[4] M. M. Maas. "Concepts in advanced AI governance: A literature review of key terms and definitions". *AI Foundations Report,* vol. 3, 2023.

[5] A. Casheekar, A. Lahiri, K. Rath, K. S. Prabhakar and K. Srinivasan. "A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions". *Computer Science Review*, vol. 52, p. 100632, 2024.

[6] A. Mary Sowjanya and K. Srividya. ividyaSrividyaSrividya24. em*Conversational Artificial Intelligence*. Wiley, New Jersey. pp. 713-725, 2024.

[7] D. Das, C. Prasad and J. Geetha. "Intelligent Conversational AI for Microsoft Teams with Actionable Insights". In: *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE. pp. 1-5, 2024.

[8] L. Jain, R. Ananthasayam, U. Gupta and R. Radha. "Comparison of rule-based chat bots with different machine learning models". *Procedia Computer Science*, vol. 259, pp. 788-798, 2025.

[9] E. Yurdakurban, K. G. Topsakal and G. S. Duran. "A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information". *Journal of Stomatology Oral and Maxillofacial Surgery*, vol. 125, no. 5, p. 101757, 2024.

[10] A. M. Saghiri, S. M. Vahidipour, M. R. Jabbarpour, M. Sookhak and A. Forestiero. "A survey of artificial intelligence challenges: Analyzing the definitions, relationships, and evolutions". *Applied Sciences*, vol. 12, no. 8, p. 4054, 2022.

[11] Y. Dong, J. Hou, N. Zhang and M. Zhang. "Research on how human intelligence, consciousness, and cognitive computing affect the development of artificial intelligence". *Complexity*, vol. 2020, pp. 1-10, 2020.

[12] I. H. Sarker. "AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems". *SN Computer Science*, vol. 3, no. 2, p. 158, 2022.

[13] A. P. Chaves, J. Egbert, T. Hocking, E. Doerry and M. A. Gerosa. "Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots". *ACM Transactions on Computer-Human Interaction*, vol. 29, no. 2, pp. 1-38, 2022.

[14] K. Mageira, D. Pittou, A. Papasalouros, K. Kotis, P. Zangogianni and A. Daradoumis. "Educational AI Chatbots for content and language integrated learning". *Applied Sciences*, vol. 12, no. 7, p. 3239, 2022.

[15] A. Bewersdorff, K. Seßler, A. Baur, E. Kasneci and C. Nerdel. "Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters". *Computers and Education: Artificial Intelligence*, vol. 5, p. 100177, 2023.

[16] N. Haristiani. "Artificial intelligence (AI) chatbot as language learning medium: An inquiry". *The Journal of Physics: Conference Series*, vol. 1387, no. 1, p. 012020, 2019.

[17] A. S. E. AbuSahyon, A. Alzyoud, O. Alshorman and B. Al-Absi. "AI-driven technology and chatbots as tools for enhancing english language learning in the context of second language acquisition:

A review study". *International Journal of Membrane Science and Technology*, vol. 10, no. 1, pp. 1209-1223, 2023.

[18] I. Beltagy, A. Cohan, R. Logan IV, S. Min and S. Singh. "Zero- and Few-Shot NLP with Pretrained Language Models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 32-37, 2022.

[19] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace and S. Singh. "Autoprompt: Eliciting knowledge from language models with automatically generated prompts". *arXiv preprint arXiv:2010.15980*, 2020.

[20] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru,… & X. L*i*. "Few-shot learning with multilingual language models," *arXiv preprint arXiv:2112.10668*, 2021. doi: 10.48550/arXiv.2112.10668

[21] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes and A. Mian. "A comprehensive overview of large language models". *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1-72, 2025.

[22] B. K. Arif and A. M. Aladdin. "A comparative analysis of ChatGPT and traditional machine learning algorithms on real-world data". *Kurdistan Journal of Applied Research*, vol. 10, no. 2, pp. 93-118, 2025.

[23] E. A. Alomari. "Unlocking the potential: A comprehensive systematic review of ChatGPT in natural language processing tasks". *Computer Modeling in Engineering and Sciences*, vol. 141, no. 1, pp. 43-85, 2024.

[24] A. Asai, S. Kudugunta, X. Yu, T. Blevins, H. Gonen, M. Reid, Y. Tsvetkov, S. Ruder and H. Hajishirzi. "BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1: Long Papers. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1771-1800, 2024.

[25] P. Efimov, L. Boytsov, E. Arslanova and P. Braslavski. "*The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer*". Springer, Berlin. pp. 51-67, 2023.

[26] E. Razumovskaia, I. Vulić and A. Korhonen. "Data augmentation and learned layer aggregation for improved multilingual language understanding in dialogue". In: *Findings of the Association for Computational Linguistics: ACL 2022*, Stroudsburg, PA, USA. pp. 2017-2033, 2022.

[27] W. Huang, K. F. Hew and L. K. Fryer., 2022.l Linguistics: ACL 2022tilingual language understanding in dialogue". In: sks". ta". uage acqguage acqe acqag*Journal of Computer Assisted Learning,* vol. 38, no. 1, pp. 237-257, 2022.

[28] G. A. Santos, G. G. de Andrade, G. R. S. Silva, F. C. M. Duarte, J. P. J. Da Costa and R. T. de Sousa. "A conversation-driven approach for chatbot management". *IEEE Access*, vol. 10, pp. 8474-8486, 2022.

[29] W. Maeng and J. Lee. "Designing a Chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ML-based Chatbot". In: *Asian CHI Symposium 2021*. ACM, New York, NY, USA. pp. 160-166, 2021.

[30] A. Sakshi, T. Mehrotra, P. Tyagi, and V. Jain, "Emerging trends in hybrid information systems modeling in artificial intelligence". In: *Hybrid Information Systems*. De Gruyter, Germany. pp. 115-152, 2024.

[31] M. Orosoo, I. Goswami, F. R. Alphonse, G. Fatma, M. Rengarajan and B. Kiran Bala. "Enhancing Natural Language Processing in Multilingual Chatbots for Cross-Cultural Communication". In: *2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE. pp. 127-133, 2024.

[32] J. Guerrero-Ibáñez, S. Zeadally and J. Contreras-Castillo. "Sensor technologies for intelligent transportation systems". *Sensors*, vol. 18, no. 4, p. 1212, 2018.

[33] H. Leunard, R. Rachmawati, B. N. Zani and K. Maharjan. "GPT Chat: Opportunities and challenges in the learning process of Arabic language in higher education". *Journal International of Lingua and Technology,* vol. 2, no. 1, p. 10, 2023.

[34] A. M. Aladdin, Y. N. Bakir and S. I. Saeed. "The effects to trend the suitable OS platform". *Journal of Advances in Natural Sciences*, vol. 5, no. 1, pp. 342-351, 2018.

[35] V. İnci Kavak, D. Evis and A. Ekinci. "The use of ChatGPT in language education". *Experimental and Applied Medical Science*, vol. 5, no. 2, pp. 72-82, 2024.

[36] F. Mosaiyebzadeh, S. Pouriyeh, R. Parizi, N. Dehbozorgi, M. Dorodchi and D. Macêdo Batista. "Exploring the Role of ChatGPT in Education: Applications and Challenges". In: *The 24th Annual Conference on Information Technology Education*. ACM, New York, NY, USA, pp. 84-89, 2023.

[37] V. Goar, N. S. Yadav and P. S. Yadav. "Conversational AI for natural language processing: An review of ChatGPT". *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 3s, pp. 109-117, 2023.

[38] G. Sharma and A. Thakur. "ChatGPT in drug discovery". *ChemRxiv*, 2023.

[39] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau and L. Chen. "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration". *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277-304, 2023.

[40] A. M. Aladdin, R. K. Muhammed, H. S. Abdulla and T. A. Rashid. "ChatGPT: Precision Answer Comparison and Evaluation Model". 2024. DOI: 10.36227/techrxiv.172833414.47483047/v1

[41] L. Kohnke, B. L. Moorhouse and D. Zou. "ChatGPT for language teaching and learning". *RELC Journal*, vol. 54, no. 2, pp. 537-550, 2023.

[42] J. S. Barrot. "Using ChatGPT for second language writing: Pitfalls and potentials". *Assessing Writing*, vol. 57, p. 100745, 2023.

[43] P. Zangrossi, M. Martini, F. Guerrini, P. De Bonis and G. Spena. "Large language model, AI and scientific research: Why ChatGPT is only the beginning". *The Journal of Neurosurgical Sciences*, vol. 68, no. 2, 2024.

[44] B. F. Gonçalves and V. Gonçalves. "*Artificial Intelligence Language Models: The Path to Development or Regression for Education*?" Springer, Berlin. pp. 56-65, 2024.

[45] D. S. M. Pereira, F. Mourão, J. C. Ribeiro, P. Costa, S. Guimarães and J. M. Pêgo. "ChatGPT as an item calibration tool: Psychometric insights in a high-stakes examination". *MedTeach*, vol. 47, no. 4, pp. 677-683, 2025.

[46] M. K. Audichya and J. R. Saini. "ChatGPT for Creative Writing and Natural Language Generation in Poetry and Prose". In: *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*. IEEE. pp. 1-7, 2023.

[47] A. Azaria. "ChatGPT: More Human-Like Than Computer-Like, but Not Necessarily in a Good Way". In: *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE.

pp. 468-473, 2023.

[48] H. Al-Khatri, N. Al-Azri, H. Hassan, R. A. Maamari and D. A. Khatri, "ChatGPT applications in active learning in higher education: A restricted systematic review," In: *Proceedings of the 11th International Conference Higher Education Advances* (HEAd), 2025. Available from: https://ocs.editorial.upv.es/index.php/HEAD/HEAd25/paper/view/20029

[49] D. O. Hasan, A. M. Aladdin, A. A. H. Amin, T. A. Rashid, Y. H. Ali, M. Al-Bahri, J. Majidpour, I. Batrancea and E. S. Masca. "Perspectives on the impact of E-Learning pre- and post-COVID-19 pandemic-the case of the Kurdistan Region of Iraq". *Sustainability*, vol. 15, no. 5, p. 4400, 2023.