Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI:
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Contextual AI Assistants for Multilingual Support: Architectures, Applications, Challenges, and Future Research Directions

Ammar Khalid
Independent Researcher, Techstrike Company
Corresponding Author: Ammar Khalid, **E-mail**: Mohummad.ammar@gmail.com

| ABSTRACT

There is a growing trend of contextual AI assistants in multilingual support for the education, health care, customer service, public administration, tourism, enterprise communication sector. Unlike scripts-driven or question-answer specific chatbots, contextual AI assistants response in a more relevant and adaptive fashion, using conversation history, user intent, domain expertise, retrieved documents, cultural signals, and task-specific, pertinent information. Transformer models that address languages, multilingual pre-trained models, large language models, and retrieval-augmented generation have all made great strides in improving the ability of AI assistants to comprehend and respond in various languages. However, support for multiple languages is not consistent, particularly for languages of low resource status, non-native languages, code-mixed language and culturally appropriate language. This review consolidates and summarizes literature on the evolution, their domains of use, methods of evaluation, and ethical challenges of multilingual language models (MLM), context-aware dialogue systems, retrieval-augmented generation, and applications of contextual AI assistants. While existing systems are becoming more effective at cross-lingual transfer, more effective ability to reason from context, and domain-grounded response generation, they have a number of limitations, primarily in the areas of hallucination, language bias, privacy, explainability, benchmark imbalance, and cultural adaptation. This paper outlines and categorizes contextual multilingual AI assistants and outlines future research directions for creating more reliable, inclusive, transparent, and human friendly AI assistance programs. The results point to the potential for a more powerful race of multilingual AI assistants in the future, integrating these features: language support, context management, retrieval grounding, culturally-appropriate interaction design, and comprehensive assessment tools and dialog systems designed for different linguistic communities.

| KEYWORDS

Contextual AI assistants; multilingual support; conversational AI; large language models; retrieval-augmented generation; context-aware dialogue systems; low-resource languages; multilingual NLP; human-computer interaction; AI ethics

| ARTICLE INFORMATION

ACCEPTED: 15 June 2026 **PUBLISHED:** 27 August 2026 **DOI:** 10.32996/jcsts.2026.8.9.1

1. Introduction

Artificial Intelligence (AI) assistants are proving to be key in connecting people and digital systems. They can be used to build question answering systems, navigation services, recommendations, translation systems, decision-making systems, and systems used for communication automation [2]. Some of the previous conversational agents were primarily based on a set of rules, keyword matching or very short-range intent classification. These systems were effective for limited uses but were not good when users switched languages, used informal text, mentioned previous turns in the conversation or queried with situations that required domain knowledge. The limitation has been addressed by transformer based natural language processing that allows flexible and context-sensitive language processing. Vaswani [38] laid the groundwork for the transformer architecture that serves as the basis of the modern language models, and Devlin [11] showed the effectiveness of the use of a contextual language representation for natural language understanding tasks using BERT.

As digital platforms address the needs of a variety of linguistic and cultural groups, the need for multilingual AI assistants has grown. Systems need to automatically-translate in response when asked in other languages in education, healthcare, public services, ecommerce and customer service, which demands systems that recognize both the individual's level of digital literacy and their spoken language [10]. But, multilingual support is not just about translation. Key elements of effective support are: understanding what users are trying to do, maintaining user context, being aware of user culture, searching for relevant information, and providing linguistically and culturally adequate answers. To direct the cross-lingual transfer various other models were developed, including multilingual models like XLM-R with unequal performance to low resource languages, dialects and code-mixed communication [9].

Understanding the context is key to intelligent AI support. A contextual assistant is an assistant that understands the previous turns that a user has made in the conversation, what the user wants, the overall purpose of the task, what the domain allows and what the user might know. An assistant in a multilingual situation may, for instance, need to maintain history of the conversation in the various languages, retrieve documents in one language and make them understandable to respondents in another. The use of few-shot learning and flexible task adaptation to learn from small amounts of data or adapt to novel tasks has enhanced these abilities, as has the ability to generate fluent utterances, using LLM models [6, 31]. More generally, open multilingual models like BLOOM [34] also highlight need for more extensive language coverage in general large-scale AI to come to fruition. However, these models can fail to produce accurate or valid information, particularly in extremely critical contexts like healthcare, immigration, law and public administration [3].

Nearly one way that retrieval-augmented generation is being used to increase factual knowledge is by using it to increase factual grounding. RAG systems do not just need to rely on knowledge that is hard coded into the system weights, they also need to be able to retrieve the relevant documents or passages which can serve as evidence for output generation. This method was found to improve the performance on knowledge-based NLP tasks by Lewis [21]. In multilingual settings, RAG can be used to answer questions in the language, domain and policy code set that they are prepared to answer, including low-resource language systems like Swahili conversational systems [26]. Retrieval errors, reliability of sources, document language mismatch, latency, and faithfulness evaluation, however, pose other problems for RAG.

In this review, underlying architectures, context management, RAG, evaluation techniques, applications, and ethical considerations of contextual AI assistants in multilingual scenarios are explored. It's most significant feature is a hierarchical organization and overview of context-oriented, multi-lingual AI helpers.

2. Methodology of the Review

To take the required steps to ensure the conservation of the protocol of a systematic review in this study, the approach of a structured narrative review was chosen. The goal is to compile available and information and group it into important themes covering contextual AI assistants and multilingual assistance. The review is organized according to transparent identification, screening, selection and synthesis as recommended in the PRISMA-Style reporting guidelines [29].

To conduct the literature search, the following keywords were used: "context-aware conversational agents," "contextual AI assistants," "multilingual support," "multilingual dialogue systems," "LLM (large language models)," "RAG (retrieval-augmented generation)," "cross-lingual transfer," "low-resource languages," and ".multilingual chatbot evaluation." Peer-reviewed papers available in journals, conference papers, influential papers in preprints and market-proven benchmark studies were given priority between the years 2017 and 2026.

This included studies that covered multilingual natural language processing, conversational AI, context-aware dialogue systems, large language models, retrieval-augmented generation, low-resource language technologies, multilingual evaluation, or ethical issues regarding the use of AI to aid in communication. Basic research was included along with technologies directly relevant to the era of modern AI assistants, like transformers, contextual language models, and large-scale pre-training. Studies involving conversations or low resource support, as well as those pertaining to multilingualism and retrieval grounding were included in recent studies.

Researches that did not include language-processing capabilities, those that concentrated on general AI, not on the language part, or those that were based on very limited rules and did not have "real" processing capabilities, and those that were paper records were duplicated or not academic opinion papers or those that did not provide enough technical details, were omitted from the study. The selected research studies fell into six categories, including multilingual language models, context-aware dialogue modelling, retrieval augmented generation, application domains, evaluation and benchmarking, and ethical issues.

Since there are many differences across various models, datasets, tasks and metrics in the field, a conceptual synthesis approach was employed as an alternative to statistical meta-analysis. For instance, intent recognition and dialogue-state tracking are assessed in some studies, retrieval quality, factuality is assessed in some studies, user satisfaction or cross-lingual transfer in some studies. Themes of comparison, elicitation of theme taxonomy and gaps in research are reflected throughout this review, because many of the dialogue benchmarks are still English oriented.

3. Conceptual Foundations of Contextual AI Assistants

Contextual AI assistants are AI-powered conversational tools that are able to offer adaptive assistance by understanding user input in relation to other surrounding aspects of a conversation, language, information and environment. Contextual assistants can leverage data from the previous turn, user intentions, domain knowledge, and external sources to deliver better responses as opposed to simple turn-based chatbots that respond based on an intent or script [12]. In use, a contextual assistant would need to be able to deduce the meaning of the user's query, the motivation behind the request, information that has already been mentioned, the source of information and how to present the answer in the user's language of preference.

Prospects of contextualization are closely related to the progress of neural language modeling. Advancements in neural language modeling are closely linked to the rise of contextual AI assistants. They adopted the transformer architecture which enabled them to model relationships between tokens that span a whole sequence of tokens by adding self-attention as a mechanism, yielding more powerful ways of accounting for long-range dependency than many previous recurrent architectures [38]. This was significant for conversational AI since a great deal of the appearance of sentences relies on long-distance interactions such as pronouns, references, ellipsis and implicit meaning in multiple turns of the conversation. In turn, BERT improved over the previous methods in a several ways, including by using deep bidirectional models and fine-tuning them for specific tasks, such as question answering and natural language inference [11]. All of this helped bring conversational systems from rule-based flows to more flexible and language driven systems.

Table 1 Summary of Key Concepts in Contextual Multilingual AI Assistants

Concept	Meaning	Relevance to the Study
Contextual AI assistant	An AI system that uses conversation history, user intent, task state, domain knowledge, and external information to generate adaptive responses.	Forms the main object of the review.
Multilingual support	The ability to understand, retrieve, translate, and generate information across multiple languages.	Central to inclusive digital communication.
Context awareness	The ability to interpret user input based on previous turns, user needs, domain information, and situational cues.	Improves coherence and relevance in multi-turn interaction.
Cross-lingual transfer	The ability of a model trained on one or more languages to perform tasks in another language.	Supports languages with limited training data.
Low-resource language	A language with limited digital corpora, annotated datasets, or NLP tools.	Highlights fairness and inclusion challenges.
Retrieval-augmented generation	A method that retrieves external evidence before generating an answer.	Reduces hallucination and improves factual grounding.
Code-switching	The use of two or more languages within the same conversation or sentence.	Common in real multilingual communication.
Human-in-the-loop AI	A design approach where human experts review, guide, or correct AI outputs.	Important in high-risk domains such as health, law, and public services.

3.1 Context Awareness in AI Assistants

Context awareness: An AI system's capability to leverage context, or surrounding information, when deciding what to interpret or about what to reply. The context in AI for conversational applications can be segmented into multiple types. Dialogue context: using previous user and/or system turns. Task context refers to the mission or task which is in a current state, such as making an appointment, going through an administrative procedure, or sales policy. User context can refer to a user's language preference, level of expertise, accessibility needs and prior interactions. Other contexts like health care, education, immigration, agriculture, customer service, have specialized knowledge that may be called domain context. Local norms, conventions of politeness, and idioms and socially appropriate means of explaining is included in cultural context. Context is retrieved as external documents knowledge base/databases used to aid the response [5].

In multi-turn interaction, giving context is particularly very important. A user can start with a general question that becomes more specific later in the incoming chat, switch to a different language halfway in the conversation, and continue the conversation with additional questions by using the vague terms like “that”, “there”, or “the second option”. With a non-contextual system, the system can construct all possible responses to each query without considering whether they are relevant or appropriate, certainly not allowing the context to influence them. Rather, a contextual assistant should maintain context and refine its grasp of the conversation as it evolves [30]. For task-oriented dialogue systems, this is typically related to dialogue-state tracking, a process where the system monitors user goals, constraints and values of the system's various slots through conversation turns. This problem has been formalized with multi-domain datasets like MultiWOZ, which include many conversations within multiple domains, including forums, restaurants, taxi and attractions. It has been formalized using multi-domain datasets, such as MultiWOZ, which is a collection of conversations over a variety of domains including forums, restaurants, taxis, and attractions [7].

With multilingualism things go a bit more complex with respect to context awareness. Different language variations can be used; existing words may be borrowed, or local variations may be used along with a dominant language; and cultural words and concepts may be employed. This is not enough to make it easy to be translated. For example, an assistant that provides public service support should not just be multilingual, capable of translating the terms, but also explain the meaning in a manner that does not differ from the social and administrative context of the user. Likewise, the healthcare assistant must ensure that the medical significance remains intact across languages but that it doesn't become oversimplified. This can cause misinformation, exclusion or unsafe advice in these contexts of failure [24].

Table 2 Types of Context Used in Contextual Multilingual AI Assistants

Type of Context	Description	Example in Multilingual Support
Dialogue context	Previous user and assistant turns in the conversation.	Understanding what “that document” refers to from an earlier message.
User context	User profile, language preference, literacy level, accessibility needs, or previous interactions.	Explaining a policy in simple Spanish for a beginner user.
Task context	The current goal or service being completed.	Helping a user complete a visa form or book an appointment.
Domain context	Specialized knowledge from a field such as healthcare, education, finance, or law.	Explaining medication instructions using verified health information.
Cultural context	Local norms, politeness conventions, idioms, and socially appropriate expressions.	Using respectful forms of address in Japanese, Arabic, or Hindi.
Situational context	Time, location, device, service channel, or environmental condition.	Giving location-specific tourism or emergency guidance.
Retrieved context	External documents, databases, FAQs, or policies retrieved before response generation.	Retrieving a university admission rule and explaining it in Urdu.
Multimodal context	Information from speech, image, video, document uploads, or forms.	Reading a government form image and explaining it in the user’s language.

3.2 Multilingual Support in Conversational AI

Multilingual support means the capability of an AI assistant to process, understand, retrieve and create information in more than one language. This can be as simple as language identification and machine translation. At higher levels, it demands the ability to understand and respond at a cross lingual level, to transfer the information, to adapt the information to different cultures, and to ensure comparable performance across language. Multilingual representation learning is integral to modern multilingual conversational AI, involving the training of models on large collections of texts in multiple languages, and enabling similar expressions across different languages to be embedded together in a shared space [32].

There are significant examples like multilingual BERT and XLM-R for this development. Both the multilingual and BERT extensions showed how to make use of multilingual corpora for representation learning to enable cross-lingual dialog transfer [27]. Both the multilingual extension and BERT exhibited the power of representation learning from a multilingual corpus for cross-lingual dialog transfer learning [11]. The paper by Conneau [9] demonstrated that XLM-R was well suited for large-scale unsupervised cross-lingual representation learning, and achieved good results across multilingual understanding tasks. The models enabled systems to be developed to be very specific in one language and to be able to transfer other capabilities to the other language(s). The extent of training data in each language and their quality, however, typically have a significant impact on performance. There tends to be a greater benefit for high resource languages (those that have many resources, such as English, Spanish, French, and Chinese) than in low resource languages [4].

Multilingual AI assistants still have a way to go before they can be of service in the low-resource language category. The availability of annotated datasets and the uniformity of spelling conventions, as well as fewer numbers of digital documents and less availability in large digital corpora, are smaller in many languages. Community-driven datasets like MasakhaNER have proven useful for African languages, and that most NLP research under representatives are not represented in research. Many of the African language communities are underrepresented in NLP research and community driven datasets like MasakhaNER have proven to be useful [16]. This is a direct focus point for multilingual support as there is a possibility of an assistant working well in high resource languages but failing in low resource languages causing perpetuation of existing digital inequalities. Therefore, assessing multilingual AI assistants in terms of their average performance is not sufficient; performance variations by language and for various user groups should also be assessed.

Table 3 Multilingual Support in Conversational AI

Model/Technology	Main Function	Strengths	Limitations
mBERT	Multilingual contextual language representation.	Supports many languages and enables cross-lingual transfer.	Uneven performance across high- and low-resource languages.
XLM-R	Large-scale cross-lingual representation learning.	Strong multilingual understanding performance.	Requires fine-tuning and may still underperform for underrepresented languages.
mT5	Multilingual text-to-text generation.	Useful for translation, summarization, QA, and generation tasks.	Computationally demanding and dependent on multilingual training quality.
BLOOM	Open multilingual large language model.	Promotes multilingual and open-access LLM development.	Performance varies by language and domain.
GPT-style LLMs	General-purpose instruction following and generation.	Strong fluency, reasoning, and few-shot adaptation.	Risk of hallucination, bias, and limited transparency.
RAG systems	Retrieve external evidence before generation.	Improves factual grounding and source-based answers.	Depends heavily on retrieval quality and source reliability.
Cross-lingual embeddings	Represent semantically similar content across languages.	Useful for multilingual search and retrieval.	May fail for dialects, code-switching, or low-resource languages.
Speech-based multilingual AI	Processes spoken input and output across languages.	Improves accessibility for low-literacy and mobile-first users.	Sensitive to accents, noise, dialects, and speech-recognition errors.

3.3 Large Language Models and Retrieval-Augmented Generation

By allowing for more much-less-tuned generation, instruction following and few-shot adaptation, LLM's transformed the design of Generative AI's assistant. Brown [6] showed that scaling model size and training data makes for good task generalization via prompting. To that end, it has been shown that many NLP tasks can be reformulated as problems involving generation of text from input data [31], such as provided by text-to-text models. New open LMs (with more languages, more transparent collaborative processes) of the type of OpenAI's GPT-3 or BLOOMs were also introduced, showing that there is the potential to build larger LLMs with more languages and more transparent processes of development [34].

But large language models also introduce the danger to provide contextual multilingual support. May produce fluent answers but answers are incorrect; may overgeneralize and use dominant language patterns; may lose the meaning of the matter in the target language; or may come up with culturally inappropriate answers. One approach that tackles some of the issues mentioned above is retrieval-augmented generation (RAG), which ties the generative part of AI models with external data stores of knowledge [13]. They then use this to restrict the answer it generates in the RAG systems so that it uses relevant information retrieved from a document collection and rates it accordingly. An advantage of this method, for multilingual support, is that the assistant can refer to relevant, and possibly corpus, up-to-date and language specific documents to find answers other than model memory [21].

RAG can be used in the context of multilinguistics, where a user asks a question in one language and RAG system returns the relevant evidence in a different language. It can also aid in local information retrieval by searching the information bases of official documents, institutional policies or the community-specific knowledge base in the localized language. However, multilingual RAG needs to be well-designed. Even if the generation model is high quality, inaccurate interpretation can result from retrieval failure, failure to properly translate, a lack of matching with documents, and/or poor source ranking. Thus, the quality of the contextual multilingual assistants should be assessed in a full-fledged system that relates to the different stages of language understanding, language retrieval, grounding, language generation, and the user experience [25].

4. Architectures and Technologies for Multilingual Contextual Assistants

The framework of a multilingual AI assistant designed for contexts generally is structured into several layers. The architecture of a contextual multilingual AI assistant usually involves several layers: input processing, language identification, natural language understanding, context management, and knowledge retrieval, response generation, safeguarding and output delivery. These functions have been implemented as a single large language model in simple systems, but in more complex systems, they are split into different components, to boost reliability, transparency, and control. The layers are particularly relevant for multilingualism, as the presence of language variance, domain-specific terminology, cultural context and code switching makes challenges in interaction go beyond what is encountered in a traditional single language scenario [28].

The first one is ‘Input Processing'. In order to assist the user, the assistant needs to be able to recognize the language the user uses, understand spelling variations, correct informal expressions or recognize when the user has typed in something in different languages or mixing languages together. In a multilingual community, often words from different languages will be used together in the same sentence, called code switching, or two languages will be spoken adjacent in the same sentences. A single language per conversation setting could be misinterpreted, as would be the case in healthcare, education and public services when users employ local terminology describing technical or official terms.

The second is the understanding of natural language. These are features such as intent recognition, entity extraction, sentiment detection, and dialogue-state tracking. Someone in the customer support team, for instance, will have to determine if the user wants some information, wants to report a problem, is asking for every payback or wants any technical guidance. There can be variability in the means of expression for the same meaning among the languages and cultures in multilingual settings. To solve this, there are methods like multilingual BERT and XLM-R, which are able to learn common representations across languages, but even they do not seem to perform as well for low-resource languages [9, 11].

The third layer is the context management. This component keeps and updates information about dialogue that is relevant to the conversation. For example, when a user inquires about requirements for admission, and proceeds to ask, “What documents do I need for that?” the Bot should relate “that” to the previous question. Context management can include continuation of the dialogue-state, user preferences, goals regarding tasks, or summarizing previous turns. It can be hard to have long calls, though, since the models come with a maximum number of contexts they can remember and they could get off topic.

The fourth layer is about knowledge access. Retrieval-Augmented Generation, which retrieves relevant documents, passages, and database or table entries before generating a response, is gradually becoming used by modern assistants when appropriate. This puts the answers in context and is helpful when accuracy is critical, like in the fields of healthcare, immigration, legal and university administration, and technical support. A new study by Lewis [21] demonstrated that retrieval-augmented generation works well when the task is related to knowledge and combines a retrieval component with generative models.

“Response Generation” is the fifth layer. Large language models can return a fluent and flexible response; however, the advantage of fluency doesn't ensure accuracy. In Multilingual Support responses should convey factual information, be sensitive to user language preference and include relevant tone. T5 models and generative pre-trained transformer (GPT) models, together with multilingual LLM models, can deliver better quality of generations, albeit with the tendency to deliver possibly hallucinatory and off-cultural responses [6, 31, 34].

The last layer is security and quality assurance, such as protecting privacy, detecting bias, filtering toxicity, signaling uncertainty and escalating to human support. Based on AI literature, contextual multilingual assistants can be roughly categorized as model-based, retrieval-based, task-based, personalized, low resource, multimodal, or agentic assistants.

Table 4 Taxonomy of Contextual Multilingual AI Assistants

Assistant Type	Description	Typical Use Case	Key Requirement
Model-centered assistant	Relies mainly on a multilingual or general-purpose language model.	General question answering and basic support.	Strong multilingual language model.
Retrieval-grounded assistant	Uses RAG to retrieve evidence before answering.	University, healthcare, legal, or policy support.	Reliable document retrieval and answer faithfulness.
Task-oriented assistant	Helps users complete a specific goal.	Booking, form filling, troubleshooting, or service navigation.	Dialogue-state tracking and intent recognition.
Personalized assistant	Adapts to user history, preferences, and expertise level.	Education, workplace support, customer service.	Secure user memory and privacy controls.
Low-resource assistant	Designed for languages with limited data and tools.	Rural advisory, indigenous language support, African and South Asian languages.	Community datasets and transfer learning.
Multimodal assistant	Combines text, speech, image, forms, or video.	Public services, tourism, accessibility, healthcare.	Strong multimodal processing and error handling.
Agentic assistant	Uses tools, APIs, planning, and multi-step reasoning.	Enterprise workflows, travel planning, institutional services.	Tool reliability, monitoring, and safety controls.

5. Application Domains

Contextual multilingual AI assistants are becoming ubiquitous across industries where users are needing real-time, language intelligent support in a timely and accessible way. Healthcare is a one key application area. Multilingual healthcare assistants are able to explain health issues, provide directions on medications, how and when to make appointments, diet advice, public health services, and health prevention tips. In public health communication, conversational AI can help facilitate widespread messaging in terms of multiple language delivery, and customization to suit varying literacy levels. In short, the healthcare sector is one of the most high-risk areas involving health assistants, which requires that they refrain from an unsupported diagnosis, maintain clinical meaning in the translated text, and make it clear where end-users are expected to reach out to a qualified professional. This is another great use for retrieval grounding because reliable and up-to-date information has to be used for medical information [23].

There's a great emphasis on education, too. Multilingual customized AI can offer tutoring, language-learning assistance, academic advice, writing feedback and institutional information. Students could gain advantages from having their explanations in their mother tongue and still applying academic vocabulary in the language (L2) of instruction in the multilingual classroom. The AI Assist can interact with university systems, course registration, rules for scholarships, administrative process etc. to make it easier for international students to navigate. Educational assistants, however, need to be carefully crafted to facilitate learning, and avoid the temptation to supplant its attainment. They should stimulate thinking and teaching and not just answer the question [35].

Another area where contextual multi-lingual assistants can play a significant role is customer or enterprise support. Many customers whose native language is not the same as yours are expecting information to be easily available and that it will be translated for them. Customers from around the world may use a different language and expect a response promptly. AI assistants can help field general common queries, address product issues, handle service requests, and escalate more complex queries to humans. This means that the customer experience will be better as the assistant will be aware of previous information, save from having to repeat information, and respond to the user's specific problem. Translation or retrieval of support in multiple languages can enhance waiting times and accessibility but can also be detrimental when they are not accurate or the translation is of poor quality [15].

Another significant domain for public administration and immigration services are highly relevant. In many cases, the form can be complicated, needs to be filled out in specific ways, have certain eligibility requirements and deadlines, and contain legal jargon and procedural information [33]. Public services can be more accessible to migrants, refugees, international students and minority language users through the use of multilingual contextual assistants. Update: This pertains to an online domain being used as well, which can have serious comes and goes if the right info is incorrect, as that can impact the legal standing, eligibility for benefits or conform to official requirements. Thus, public administrators' assistants needs to be retrieval-centered, source-attributed clear, and escalation to human administrators uncertain cases [14].

Multilingual contextual support is also used to support tourism and cultural services. Using AI assistants, they can offer suggestions that are dependent on location, translate languages for them, plan their itinerary, explain the backstory, and provide advice on emergency situations. In this area, cultural context can be particularly relevant; users can pose questions regarding socially appropriateness, etiquette, local laws or culturally sensitive areas. The role of an assistant isn't just to convey information on travel, but to convey it in a manner that encourages respectful and informed interactions [40].

Multilingual contextual assistants are also an emerging field in the area of agriculture and rural advisory services. There are times when farmers require information that is on time such as crop disease, weather, utilization of fertilizers, pest management, market prices and government schemes. This information needs to be given in local languages/dialects in many areas. A contextual assistant will be able to interpret the question posed by the user, search and find the relevant agricultural information and offer concrete and user-friendly suggestions. But rural and low-resource areas pose technical challenges including lack of connectivity, dialect variation, low digital literacy and quality of local language data. Therefore, for this reason, agricultural assistants can be optimized to deal with low bandwidth intervention, and mechanisms for uncertainty-handling and human experts should be built [17].

Multilingual contextual AI is also needed in the financial industry and in supporting employees. Multilingual contextual AI is also growing in demand across the financial sector and in providing support for employees at the workplace. In many situations, banks, insurance companies and human-resource departments need to translate complicated policies to a different language for their users. Support staff can describe claim and account policies and procedures, workplace benefits, employment policies and compliance. In such scenarios, it is crucial to have an understanding of context as the assistant needs to adjust the message according to the customer's profile, area, profession, or service history. In the case of sensitive personal information, there are also concerns that financial and work place assistants will breach this privacy and, in doing so, provide more advice than they are authorized to do so [18].

But in all these areas, contextual multi-lingual AI assistants are going to rely on more than their linguistic abilities. An effective assistant will be precise, comprehensible, culturally competent, and minimize its limits in neutrality. Assistants serving in domain specifics should be linked to a trusted source of information, and should not output information that they are not certain about as facts. In high risk environments, there still needs to be a human presence. Thus, the best opportunities may not just be ones that facilitate communication with humans, but ones that are powered by AI scalability and executed in a way that is human-centered, evidence grounded and responsibly escalated [36].

Table 5 Application Domains of Contextual Multilingual AI Assistants

Domain	Example Application	Benefits	Risks
Healthcare	Symptom guidance, medication explanation, public health communication.	Improves access to health information across languages.	Risk of unsafe medical advice or mistranslation.
Education	Tutoring, academic advising, language learning, institutional support.	Supports international and multilingual learners.	May encourage overreliance or academic misuse.
Customer service	Product support, complaint handling, FAQs, service routing.	Reduces waiting time and improves global service access.	Poor responses may reduce trust.
Public administration	Forms, eligibility rules, immigration guidance, public services.	Improves access for migrants and minority-language users.	Incorrect information may have legal consequences.
Tourism	Travel guidance, cultural explanation, emergency help.	Provides real-time multilingual assistance.	May misrepresent cultural or legal information.
Agriculture	Crop advice, pest control, weather, market information.	Supports rural and low-resource communities.	Requires local knowledge and low-bandwidth design.
Finance	Banking, insurance, claims, compliance explanation.	Helps users understand complex procedures.	Privacy and regulatory risks.
Workplace support	HR policies, onboarding, technical support.	Supports diverse employees and global teams.	Sensitive employee data must be protected.

6. Evaluation Metrics and Benchmarking Approaches

Measuring conversational quality in contextual multilingual AI support is a difficult challenge, as it requires capturing multiple factors related to understanding, relevance, coherence, factuality, safety, cultural appropriateness, and user satisfaction [39]. A system can give good answers and bad information or a system can find relevant documents but be unable to describe them well to the user. Thus, assessment needs to be multi-faceted, and should not rely solely on a single measure.

Natural language understanding is an important aspect of this. This encompasses the classification of the intents, entity extraction accuracy, accuracy of filling in slots, and dialog state tracking. Such metrics can be helpful for task-oriented assistants which need to learn about the user's goals and gather information over multiple turns. The impactful datasets have been used for benchmarking dialogue systems because they have implicit multi-domain, multi-turn conversations included, including datasets like MultiWOZ [7].

The second area developed is the quality of response generation. Conventional evaluation metrics like BLEU, ROUGE and METEOR match the candidate answers with any of the reference answers, however it's restricted because conversational question can have numerous proper answers. To overcome this problem, Zhang [41] developed Semantic metrics like BERTScore that are based on the contextual embedding of words. Other parameters such as meaning preservation, grammaticality, naturalness, and cultural appropriateness should also be taken into account in multilingual situations [37].

A third area is that of context relevance. Contextual assistants need to tap into the right prior context but not get lost in irrelevant conversations. Evaluation should consider if the system uses a referential strategy properly, maintains the topic, retains the constraints of the user and does not contradict from one turn to the next. It is crucial in longer interactions where models might not recall or repeat earlier information or focus more on what has happened more recently.

Retrieval augmented assistants – retrieval quality assessment needed. The system checks on the Recall at a set value (k), on the precision at a set value (k) and on the mean reciprocal rank as well as on the normalized discounted cumulative gain. For multilingual RAG, it is important to evaluate across languages, for example by asking a query in some source language, while the answers are expressed in another source language [19].

Other traits are also vital: factuality and faithfulness. Factuality is about correctness of the answer, not faithfulness that is the evidence retrieved to support the answer. There are cases that RAG-based systems can make unsupported claims when the retrieved context is not included in the system (as missing information) or partial evidence is overgeneralized as knowledge. Thus, faithfulness of the answer, context precision, context recall and citation accuracy, as significant factors, are especially relevant in healthcare, law, finance and public administration.

Fairness in multilingualism should be measured on a by language (dialect, script, code switched input) basis. There may not be strong results in low-resource languages represented in the average score. Since the evaluation of modern multilingual conversational assistants is not only limited to translation, benchmarks like FLORES-101 are used for evaluating in multiple languages [20].

Ultimately, a human-centered usability/safety evaluation should be included. Clarity, politeness, helpfulness, trust and cultural appropriateness will be judged by native speakers and community representatives. These properties include a potential for bias, toxicity issues, privacy concerns, harmful advice, uncertainty warnings, and escalation to human experts and should be considered in safety testing.

Table 6 Evaluation Metrics for Contextual Multilingual AI Assistants

Evaluation Dimension	**Metrics/Methods**	**Purpose**
Language understanding	Intent accuracy, slot-filling F1, entity recognition F1.	Measures whether the system understands user input.
Dialogue management	Dialogue-state accuracy, task success rate, turn-level coherence.	Measures multi-turn context handling.
Generation quality	BLEU, ROUGE, METEOR, BERTScore, human fluency ratings.	Measures linguistic quality of responses.
Retrieval quality	Recall@k, precision@k, MRR, nDCG.	Measures whether relevant documents are retrieved.
Factuality	Fact-checking, hallucination rate, claim verification.	Measures whether responses are correct.
Faithfulness	Context precision, context recall, answer-grounding score.	Measures whether answers are supported by retrieved evidence.
Multilingual fairness	Performance by language, dialect, script, and code-switched input.	Identifies language-based performance gaps.
User experience	Satisfaction, helpfulness, trust, usability, response clarity.	Measures practical usefulness for real users.
Safety	Toxicity detection, bias testing, privacy audit, escalation accuracy.	Measures responsible and safe deployment.

7. Challenges and Limitations

Although there's been significant advancements, there are several big issues with multilingual support contextual AI assistants still face. The first in a series of inequalities is language inequality. High resource languages are typically the ones that are most important in the development of most high performing language models, which then tend to have lower performance on low resource languages, dialects, minority language, and oral languages that have less written data. This imbalance can further exacerbate the narrow digital divide e.g. better AI service to users who already have a strong online presence in their language.

The second problem is code switching and dialect variation. A good number of users use some mixture of languages, or actually write informally in the language/s they are communicating in digital communications, such as email, Facebook, and other messaging platforms. Multilingual models might not be well-suited to these patterns, as examples from their training data are often written examples, rather than natural speech. This serves to curtail the importance of AI helpers in genuine multilingual communities.

The third hindrance is, as mentioned, hallucination. On less well-known and topical, or narrowly focused questions, large language models can respond with assurance that doesn't align with what is actually true. Hallucination can be more difficult to identify in multilingual scenarios as evaluators might not understand all of the languages that are supported. This can be mitigated with retrieval-augmented generation by ensuring that answers are computable based on the retrieved documents, but errors in the retrieved document when it is irrelevant or otherwise incorrect or obsolete will still cause errors in answers.

Context drift is another challenge. The assistants are responsible for determining the previous information that is relevant in the multi-turn dialogues. Consider using too little context, important details may get lost; using too much, the model might get diverted. It is particularly challenging when users change Languages, Topics or levels of detail.

The quality of retrieval is also very critical for multilingual RAG systems. When the relevant document is written in a language other than the users, a user can ask a question in one language. The end result may be incorrect, even if the generation model is good, if there is insufficient cross lingual retrieval. This constitutes a bigger issue for low resource languages with limited digital knowledge bases.

Ancient adaptation of culture is also vital. Language is social, it is polite, it is authoritarian, it is funny and it also reflects its local meaning. Literal translation can yield results which are grammatically correct but socially inappropriate. But, besides being linguistically accurate, trustworthy multilingual AI should also show an understanding of cultural and emotional nuances and be designed responsibly.

Privacy and data governance is also at the forefront. Contextual assistants can analyze user records or documents, preferences, location or private information of a personal nature. With reliance on external translation tools, cloud-based models, or third party APIs, these risks are worsened. Therefore strong consent, minimization of data, anonymization and memory under control of the user are required [22].

Last but not least, issues of how to evaluate, be explainable and overly dependent are present. Numerous, if not most, benchmarks are strongly oriented toward English and don't measure the overall ability to converse in other languages. Users should be aware that the answer might be clearly grounded based on evidence or uncertain. A high level of verbal fluency can lead to a sense of authority or expertise, but assistants should be mindful of such inaccuracies, remind users to check such information, and bring any cases of risk confrontations to the attention of qualified humans.

8. Future Research Directions

Further work would benefit an inclusive, reliable, accountable and culturally aware contextual AI assistant for multilingual situations. The creation of enhanced resources for low-resource languages – annotated data sets, multilingual dialogue corpora, speech data sets, terminology banks and culturally-sound evaluation sets – is a key priority. Participation is important as you want local speakers to make sure what you have within your data sets relates to the real language, not yours or one's imposed on the group. The case study of MasakhaNER, L1 – L2 data development between the L1 community and the L2 community demonstrates the benefits of collaborative data development for underrepresented languages [1].

Another noteworthy area could be enhanced multi-language context management. Assistants need to be able to have coherent conversations over extended periods and multiple language changes, and when a user's objectives shift. This entails more good memory mechanisms, resolving references, summarizing conversations, and selecting contexts. The assistant or answer machine should be aware of which previous information is relevant and to which information to store and utilize.

This requires a culturally sensitive response generation as well. While multilingual aid is frequently perceived as primarily translation, there is a need to be sensitive to local norms, politeness conventions and institutional meanings, as well as to users' expectations, in order to be able to offer good helping. Future systems should offer adapting explanations in a non-stereotypical way at the user level, whether that implies the ability to explain things in the user's native language or adhering to his or her paradigm.

Another important research need is to achieve more reliable multi-lingual retrieval-augmented generation. These are some of the aspects that are needed to be improved in future RAG systems: Cross-Lingual Retrieval, Document Ranking, Evidence Verification, Citation Quality and Answer Faithfulness. In contexts where stakes are high (for example, healthcare, law, finance, public administration), the difference between information that is fully supported by the sources and content giving general explanations needs to be made clear.

It is also recommended to further boost multi-lingual safety and bias assessment. Culturally specific harmful expressions, slang, script or dialect in addition to different languages must be covered in safety testing. Additionally, strategies for preserving privacy like privacy-preserving approaches like on-device processing, federated learning, secure retrieval, anonymized logs, and user-controlled memory might be investigated, particularly for assistants dealing with delicate data.

It is still important to incorporate human-in-the-loop design. Multilingual AI assistants should be here to augment experts, not replace them; to take care of simple tasks and hand over to qualified humans when there's uncertainty and potential risks to a situation. Evaluation frameworks need to also become more standardized, whereby the results in terms of performance are reported individually for each language, domain and user group. Last, but not least, future work would consider the investigation of multimodal multilingual assistants, capable of integrating speech, text, images and the understanding of documents, to enhance accessibility in the mobile-first context and/or in low-literacy situations.

9. Conclusion

Multilingual contextual AI assistants are a significant advancement in the field of AI, NLP, and human-computer interaction. They integrate language understanding, dialogue history, user context, domain knowledge, retrieval grounding, and adaptive response generation to enhance the traditional chatbot systems to include these functionalities. Increased cross-linguistic and cross-disciplinary performance capabilities of AI assistants have been achieved through transformer models, BERT, XLM-R, T5, and GPT-like models, as well as open multi-lingual LLMs, and retrieval-augmented generation [8, 6, 11, 31, 34, 38].

In general, there is still uneven support for multi-language. There are many systems that work better in high resource languages than low resource languages, dialects and code-mixed communication. Context awareness is also challenging because assistants need to maintain a greater sense of context across turns, prevent context drift and also allow for embedding meaning across linguistic and cultural liminalities. RAG can increase factual grounding but will be effective if the retrieval is done well, there are reliable documents and evaluation of faithfulness.

The need of full-scale evaluation from the aspects of the understanding of the intent, coherence of the dialogue, its retrieval quality, factualness, faithfulness, multilingual fairness, cultural appropriateness, privacy, safety, and user satisfaction are pointed out here. The major part of its codes are a taxonomy of contextual multilingual AI assistants, ranging from model-centered systems to retrieval-grounded, task-oriented, personalized, low-resource, multimodal, and agentic AI assistants. These are some of the key areas for the next-generation assistant to focus on: covering for inclusive language, managing contexts, evidence-based generation, being respectful of privacy, being aware of culture, and being transparent about evaluation. The purpose of building systems that can talk many languages is to be able to support the user, accurately, respectfully, and fairly.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambua, A., Akinode, V., Mbatia, P., Manyumwa, T., Awokoya, A., Omidiora, E., Omotoso, O., Asubiaro, T., Mbonu, C., Akinyele, K., Alabi, O., Otifile, B., Onwuegbuzia, E., Oboirien, M., Ezeani, C., & Ogueji, K. (2021). MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9, 1116–1131. https://doi.org/10.1162/tacl_a_00416

[2] Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., & Le, Q. V. (2020). Towards a human-like open-domain chatbot. arXiv. https://arxiv.org/abs/2001.09977

[3] Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7, 597–610. https://doi.org/10.1162/tacl_a_00288

[4] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922

[5] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., et al. (2021). On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://arxiv.org/abs/2005.14165

[7] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (2018). MultiWOZ: A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5016–5026). https://doi.org/10.18653/v1/D18-1547

[8] Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2475–2485). https://doi.org/10.18653/v1/D18-1269

[9] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). https://doi.org/10.18653/v1/2020.acl-main.747

[10] Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia-Gonzalez, G., Hansanti, P., Hoffman, J., Jarrett, S., et al. (2022). No language left behind: Scaling human-centered machine translation. arXiv. https://arxiv.org/abs/2207.04672

[11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[12] Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. Interactions, 24(4), 38–42. https://doi.org/10.1145/3085558

[13] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv. https://arxiv.org/abs/2312.10997

[14] Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10, 522–538. https://doi.org/10.1162/tacl_a_00474

[15] Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., & Vulić, I. (2020). ConveRT: Efficient and accurate conversational representations from transformers. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2161–2174). https://doi.org/10.18653/v1/2020.findings-emnlp.196

[16] Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In Proceedings of the 37th International Conference on Machine Learning (pp. 4411–4421). https://proceedings.mlr.press/v119/hu20b.html

[17] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

[18] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282–6293). https://doi.org/10.18653/v1/2020.acl-main.560

[19] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781). https://doi.org/10.18653/v1/2020.emnlp-main.550

[20] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[21] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://arxiv.org/abs/2005.11401

[22] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out (pp. 74–81). https://aclanthology.org/W04-1013/

[23] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726–742. https://doi.org/10.1162/tacl_a_00343

[24] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607

[25] Mehri, S., & Eskenazi, M. (2020). USR: An unsupervised and reference-free evaluation metric for dialog generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 681–707). https://doi.org/10.18653/v1/2020.acl-main.64

[26] Ndimbo, E. V., Luo, Q., Fernando, G. C., Yang, X., & Wang, B. (2025). Leveraging retrieval-augmented generation for Swahili language conversation systems. Applied Sciences, 15(2), Article 524. https://doi.org/10.3390/app15020524

[27] Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., et al. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2144–2160). https://doi.org/10.18653/v1/2020.findings-emnlp.195

[28] OpenAI. (2023). GPT-4 technical report. arXiv. https://arxiv.org/abs/2303.08774

[29] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

[30] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). https://doi.org/10.3115/1073083.1073135

[31] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67. https://www.jmlr.org/papers/v21/20-074.html

[32] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., & Weston, J. (2021). Recipes for building an open-domain chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 300–325). https://doi.org/10.18653/v1/2021.eacl-main.24

[33] Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., & Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10215–10245). https://doi.org/10.18653/v1/2021.emnlp-main.802

[34] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., et al. (2022). BLOOM: A 176B-parameter open-access multilingual language model. arXiv. https://arxiv.org/abs/2211.05100

[35] Shum, H.-Y., He, X.-D., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering, 19, 10–26. https://doi.org/10.1631/FITEE.1700826

[36] Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[37] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., et al. (2022). LaMDA: Language models for dialog applications. arXiv. https://arxiv.org/abs/2201.08239

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://arxiv.org/abs/1706.03762

[39] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., et al. (2022). Ethical and social risks of harm from language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229). https://doi.org/10.1145/3531146.3533088

[40] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of NAACL-HLT 2021 (pp. 483–498). https://doi.org/10.18653/v1/2021.naacl-main.41

[41] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations. https://openreview.net/forum?id=SkeHuCVFDr