Computer Fraud and Security
ISSN (online): 1873-7056

Fine-Tuning Large Language Models for Saudi Arabic Voice Agents

Corresponding Author: Mahmoud Abdelhadi Mahmoud Safia

Corresponding Email: msafia@jccs.com.sa

Bachelor's degree in computer engineering, Aleppo University

Abstract

To support the growing voice-focused technologies in Saudi Arabia, such as innovative city solutions, government services, healthcare, and finance that require voice-assisted search and navigation, there is a need to create voice agents that will provide linguistic and cultural accuracy with high linguistic clarity and understanding of the Saudi Arabian culture. Historical systems of natural language processing (NLP) that are usually trained over Modern Standard Arabic (MSA) or generalized (dialect) corpus in general were not necessarily capable of representing the regional, phonetic, and pragmatic peculiarities of Saudi Arabic across quite different Najdi, Hijazi, Gulf, and Southern dialects. This paper discusses fine-tuning the large language models (LLMs) to enable productive spoken-dialog systems tailored to the Saudi users.

It incorporates intensive data collections with the help of multiple Saudi Arabic sources, labeling data by dialect, preprocessing the acoustic features, and fine-tuning (several stages) of transformer-based systems. The method entails the hybrid training with textual and audio data, and the performance assessment is carried out using both automatic measures (e.g., WER, BLEU) and human expertise of the trustworthiness, fluency, and sociocultural compatibility. The practical result shows that fine-tuned models can bring a far greater accuracy than baseline MSA or generic Arabic models in particular domains of use like e-government services, travel agencies specializing in religion, and triaging healthcare systems. Issues such as ethics and practicality of fairness of dialect representation, privacy of voice data, and sociolinguistic bias are crucial ethical and practical issues that the author discusses in the paper. In addition to being usable, voice agents would require cultural competence to make them inclusive and digitally equitable.

This work presents a language-aware framework to support regional language learning options in Saudi Arabian, which further provides a blueprint that can be scaled to localize the use of LLMs in less-representative linguistic contexts, and a portion of the Saudi Arabian government is likely to meet its overall AI and digital transformation vision as outlined in Vision 2030.

Keywords: Saudi Arabic, voice agents, large language models, fine-tuning, dialect adaptation

1. Introduction

1.1 Motivation and Relevance

The rapidly growing population of voices of AI-based agents, virtual assistants, and aids, e.g., Siri and Alexa, cross-functional industry-specific chatbots have revolutionized millions of people worldwide in their interaction with computers. Voice technology in the Arab world, especially Saudi Arabia, is gaining increasing adoption in e-commerce, government services, healthcare, and home automation (Alhumoud et al., 2022). The current drive of Saudi Arabia to transform digitally means that its vision of 2030 puts a lot of emphasis on artificial intelligence being a significant factor in national competitiveness as well as innovation in the sector (Saudi Data & AI Authority [SDAIA], 2021).

However, most current voice agents used in the business market are based on Modern Standard Arabic (MSA), an official form of Arabic mainly used in pop culture and education but hardly spoken across real-life contexts. Furthermore, the pervasiveness of MSA within the natural language processing (NLP) systems raises serious usability gaps for the end-users in Saudi Arabia who communicate mainly in the local regional dialects (Habash, 2010). The lack of alignment creates fewer levels of comprehension and emotions, which equate to reduced levels of reliability, satisfaction with its usage, and use of the application, particularly with finer conversational skills and applications.

The multitude of mutually intelligible Saudi Arabic dialects, which have to be considered a linguistic and computational challenge, is why Saudi Arabic can be said to present both a linguistic and computational challenge. Unlike MSA, Saudi dialects have many colloquialisms and borrowings (especially English), sociolinguistic, and culturally based pragmatics

(Al-Twairesh et al., 2018). Devoid of such dialectal support, the large language models (LLMs) risk generating unnatural, inaccurate, or socially unacceptable responses, especially in voice-based, real-time interactions.

NLP scientists in the international community have made essential progress on a low-resource or morphology-rich language conversion of an LLM. However, Arabic dialects, specifically Saudi Arabic, are underrepresented in the vast data and massive pretraining pipeline. Recent high-tech models like AraGPT2, AraBERT, or CAMeLBERT either are MSA-specific or too generic to be able to work on conversations in Saudi Arabic (Antoun et al., 2021). This aspect of the gap in related literature manifests itself in the attempt made by the current study to develop and refine language models that demonstrate integrity as applied to the linguistic, cultural, and communicative patterns of Saudi users.

It will not simply be an expression enhancement but also a need for strategic reasons. Retaining linguistic inclusiveness and cultural contextualization is essential because AI and innovative government systems are still developing, and Saudi Arabia invests in localizing them. Voice agents with dialect-aware intelligence will increase digital inclusion, enhance the quality of user interactions, and promote national objectives of AI sovereignty and reform of national public services.

1.2 The Saudi Arabic Challenge

There is no standard dialect of the Saudi Arabic language, just like the variant that existed, but this is a continuum of the regional diversities in the Kingdom. There can be four major dialectical sets, which usually exist: the Northern Arabic (Najdi), the Port of Jeddah and western corridor (Hijazi Arabic), the eastern Province (Gulf Arabic), and the southwest areas, including Asir and Najran (Southern Arabic). They are represented by some significant variation in phonology, morphology, syntax, and lexicon (Ingham, 1994; Alotaibi, 2022).

Indicatively, the Najdi dialect will sound and lack final vowels. Pronouns will have other systems than Hijazi, which is phonetically closer to MSA but with unique intonation and borrowings. In contrast, Gulf Arabic is highly affected by the surrounding Gulf regions and Persian; however, the Southern forms are highly stylized and can be challenging to understand for other areas (Prochazka & Albirini, 2020). This sort of linguistic diversity complicates the extraction of generalizable modeling and necessitates regionalized fine-tuning models.

Further, Saudi Arabia is a code-switched variety where English phrases and technical terms are easily interposed, especially by youthful and urban communities. It is common to use Arabic function words with English content terms; thus, we can find things like Shaghal the AC (turn on the AC) or Ma yhimni the deadline. It is not only lexical but functional at the same time because it shows identity, social status, and educational background (Albirini, 2016). The LLMs do not learn such patterns during training on a monolingual corpus and real-world and multi-register discourses; they perform dismally.

The next, fatal level of complexity is pragmatics, namely politeness, speech acts, and gendered discourse. An example here could be the instructions such as: *ifrah el bab (open the door), which could entail softening of some to speak in its vaguer form based on the age, gender, and social stature of the interlocutor that conventionally are encoded through Saudi Arabic but not in the general-purpose models (Ryding, 2005). Likewise, speech genres like diminutives, honorifics, etc., carry powerful cultural weight that AI agents should produce properly to avoid pragmatic failure or even offense.

In addition, gender morphology and word use are also employed, such as second-person conjugation of verbs and use of pronouns. When a voice assistant cannot distinguish between anta (you, masc.) and anti (you, fem.), it causes the users to respond negatively, or it may contradict societal rules. Such nuances require strong fine-tuning that keeps in mind variations of sociolinguistics and cultural adequacy.

1.3 Research Aims and Contributions

The research goal is to bridge the gap between the state-of-the-art architectures of LLM (current developments) and the realities of the language of Saudi Arabia (lingual facts) by systematically developing, optimizing, and testing voice-oriented language models that would align with the Saudi context. The three significant contributions of the research are as follows:

(1) Dialect-Specific LLM Fine-Tuning:

We represent fine-tuning the transformer-based LLMs, e.g., GPT, T5, or BLOOM variant, selectively trained on the Saudi Arabic corpora to fall into the Great Arabic language context, as it is about all the dialects and registers. To achieve this, in the fine-tuning stage, in addition to supervised learning (when the training data can be annotated), instruction tuning would also be exploited to align the generation output with conversational norms.

(2) Construction of a Domain-Relevant Dialectal Corpus:

The existing Arabic corpora lack the regionally marked information, which has a voice-friendly aspect and is subject to bias in favor of MSA. To overcome this gap, the study will acquire a colossal, annotated corpus encoded in dialect with Saudi social media, podcasts, call center scripts, and television talk. The emphasis will be on preserving prosody, the speaker's intent, and pragmatic clues.

(3) Multidimensional Evaluation Framework:

A test pipeline exceeding BLEU or perplexity will be proposed, where there will be human-in-the-loop, with fluency, dialectal authenticity, politeness, gender appropriateness, and cultural congruence scored alongside it. A voice agent with emotional intelligence and socially sensitive situational interactions requires a qualitative dimension.

The paper can be described as a tangential addition to the discourse on localizing both AI and paradigms of low-resource language analytics and ethically aligned design of voice interactions by demonstrating that the outcome of localizing a whole, a result realized in itself, is not accurately captured in the process of locally developing technical linguistic agents, the implementation of which embodies such a particular result, and thus in their construction (here, large language models). It further supports the strategy to have AI inclusive and context-sensitive, as that is the most significant need in multilingual communities that are fast digitizing around the globe.

Finally, the project aims to invoke the emergence of Saudi-centric voice agents who can work without regional, tongue, or social friction. Not only will these agents enhance digital service availability, but they will also create more trust and user adoption of AI technologies in Saudi-based users, moving closer to the benefits of intelligent automation.

2. Saudi Arabic Linguistics and Related Work

2.1 Linguistic Features of Saudi Arabic

Saudi Arabia is a sophisticated and regionally diversified dialect continuum encompassing the larger umbrella of Arabic occurrences. It stands out relative to Modern Standard Arabic (MSA) in a couple of linguistic features, such as infrequent phonological patterns, morphological-syntactic forms, and vocabulary that are limited linguistically.

In phonology, Saudi Arabic has plenty of emphatic consonants (e.g. / 2, / / / / 2) that produce an effect of articulation on a later vowel, which was pharyngealized and backed. Another one that stands out is vowel lengthening, and that can distinguish lexical meaning, and is phonemically contrastive, e.g.,/saba/ (he swam) /sabaa/ (he cursed) (Ingham, 1994).

Some of these Saudi dialects are morphologically and syntactically very different from MSA's. The subject-verb agreement may be quite disparate since in oral forms of language, the order of verb-subject (VS) often prevails. In contrast to MSA, most Saudi dialects have reduced inflection in the noun case suffixes, and morphological inflection is less. One such case is that the genitive case tends to be replaced with analytic possessive particle structures (e.g., 하qq, mal) instead of inflexion of the genitive case.

High levels of variation are especially encountered with lexical items and idiomatic ways of communicating in Saudi Arabia, as the dialects of Najd, Hijaz, and South Arabia, as well as the Gulf, have their own vocabulary and idiomatic ways of communicating. As an example, the term that translates as now can vary in enunciation to alhhn in the Gulf dialect, hn in Najdi, and dilhhn in Hijazi, and such a variation can indicate not only geographical but also social and cultural belonging (Alahmary et al., 2022). These local variations are not displayed effectively in generalized Arabic NLP corpora, which may obstruct specific dialectal language modeling accuracy.

Dialect	Region	Key Phonological Traits	Lexical Examples	Common Code-Switch Patterns
Najdi	Central (Riyadh, Qassim)	Final vowel dropping, affrication	"yibgha" (wants)	Shaghal the AC

Hijazi	West (Jeddah, Mecca)	Intonation closer to MSA	"dilhhin" (now)	Hybrid phrases
Gulf	East (Dammam, Khobar)	Influence of Persian, stress on final syllables	"alhhin" (now)	Mixed tech terms
Southern	South (Asir, Najran)	Stylized morphology, less mutual intelligibility	Unique idioms	Less common

Table 1: Summary of Saudi Arabic Dialect Features

2.2 Existing Work on Arabic NLP and Dialect AI

NLP in Arabic has recently shifted into the thrust with progress in recognizing a sufficient degree of difficulty in Arabic dialect, especially Saudi Arabic, in the past decade. More recently, Arabic variants of BERT are AraBERT (Antoun et al., 2020), CAMeLBERT (Inoue et al., 2021), and MARBERT (Abdul-Mageed et al., 2021), and they have pretrained BERT with a large volume of Arabic news, Wikipedia, and Arabic dialect-based social media. Surprisingly, MARBERT incorporated dialectal Arabic tweets, and it did not have performance slackness in later stages of dialect thickness.

The models, however, generalize across dialects and do not perform on higher levels of the regional accents, e.g., Saudi accents. In particular, a decreased precision applies to the MARBERT case concerning the classification of variants closely related to the Gulf variant, or variants with morphologically ambiguous Saudi articulations (Abdul-Mageed et al., 2021). Further, the Arabic GPT-like prediction models, such as GPT-Arabic or Noor GPT, are overwhelmingly trained on MSA and literary Arabic, thus lacking a natural feel for conversation in Saudi dialects.

Evaluation work has shown that the general-purpose LLMs trained on MSA are underperforming on the other dialectal analysis tasks regarding intent classification, NER, and machine translation. Elaraby and Abdul-Mageed (2021) mention that the performance in this type of dialect slows down when it struggles with morphologically richer or syntactically less-typified input unless morpho-phonological fine-tuning occurs on a dialogue level.

Inaccessibility of spoken dialogue systems is even more restricted. The currently existent Arabic conversational agents are either deterministic and rely on rule-based systems or are trained on an English-to-Arabic translation dataset, which lacks the fluidity of spoken Saudi Arabic and its informality (Zaghouani et al., 2020). Additionally, such code switching often occurs between dialectal Arabic and English, which is pervasive in Saudi speech situations (particularly in tech, retail, and healthcare sectors), and causes extra issues that are poorly tackled by most systems.

Consequently, there is still an overwhelming research opportunity regarding voice-first and dialect-sensitive language models, where the responses can be linguistically suitable and context-sensitive, especially regarding real-time verbal interactions with Saudi Arabs.

2.3 Theoretical Foundations of Fine-Tuning for Low-Resource Languages

Transfer learning: Pre trained language models. The transfer learning of transferring knowledge regarding one thing (e.g., English, MSA) to the other (underrepresented) is applied to justify fine-tuning pretrained models to low-resource languages or dialects. The transfer learning allows the models to use the pre-trained weights and adapt to the statistical characteristics of the target dialect so that the need for data and computation time is significantly lowered when adapting a given domain (Ruder et al., 2019).

Most recently, dialectal Arabic has been particularly promising with the advance of the vocabulary primitive (or adapter modules) (Houlsby et al., 2019). They are thin trainable layers that can be added to the frozen pre-trained model to perform the parameter retraining to achieve dialect-level specifications without iterative parameter retaining. This qualifies them to be used in resource-limited fine-tuning, having domain-related data that is limited or scarce in Saudi Arabia.

The encouraging future is also the few-shot learning methods, especially in the solutions based on speech, in the Saudi dialect, where labeled data is costly to create. Remarkably, only a small amount of annotated data through meta-learning or prompt tuning can grant models the power to perform astounding dialectical responsibilities without traversing the information (Brown et al., 2020). In the real world, this may translate to adjusting a basic Arabic LLM on a handful of

Saudi Arabic conversations so that it can have task-oriented comprehension, such as medical screening questions or customer service interactions.

Another essential characteristic is that it is often not taken into account that the consciousness of a sociolinguistic basis of simple dialect-language patterns. Saudi Arabia is not just a language system; it prises the elements of the social registers, gender regulations, and geographical indications of identity. The result of skipping these dimensions is decontextualized language models capable of providing tone- deaf or inappropriate responses to humorous situations in conversation. They have raised the alarm that it is essential to have context-sensitive embeddings, which incorporate user metadata, speaker demographics, or regional dialect labels to add context to the model representation to generate culturally attuned output (Blodgett et al., 2020).

Therefore, fine-tuning of Saudi Arabic voice agents should no longer rely on token level adjustment and shift towards multidimensional modelling that includes linguistic, social and pragmatic dimensions so that the model is not merely accurate but also food-for-thought.

3. Methodology: Data, Models, and Fine-Tuning

3.1 Dataset Construction and Preprocessing

3.1.1 Data Sources

The success of the Saudi Arabic voice agents hinges on the quality, variety, and detail of foundation data. A multi-modal corpus has been created in the framework of this research project and compiled based on a combination of social media communication (Twitter/X, TikTok, Snapchat), YouTube interviews, podcasts with Saudi speakers, and anonymized call center stores at the communication levels of the public sector. These sources present a rich example of the spoken Saudi Arabic (formal and colloquial varieties, presence of contextual language switching, and pragmatic discourse markers that are of primary importance to conveying the natural conversation modeling) (Habash et al., 2022; Mubarak et al., 2021).

Data Source	Dialect(s) Represented	Volume (Hours/Text Units)	Annotation Method	Notes
Twitter/X	All (esp. Najdi, Hijazi)	120,000 tweets	Manual + Automatic	High code-switch presence
YouTube Interviews	Najdi, Gulf	300 hrs	Manual transcription	High prosodic variation
Call Center Logs	Hijazi, Southern	200 hrs	Forced Alignment + Manual Review	Anonymized
Podcasts	Gulf, Southern	150 hrs	Manual	Emotionally expressive
Government Services Data	MSA, Najdi	100 hrs	Semi-supervised	Formal register mix

Table 2: Dataset Composition by Source and Dialect

3.1.2 Dialect Labeling.

Given the internal diversities of the Gulf Arabic dialects, especially in Saudi Arabia (Najdi, Hijazi, Southern, Eastern), the labelling of dialects is an essential aspect. It has been decided to choose the technique of the two-level annotation of the dialects:

- Manual Labeling: A stratified sample of data was found, with regional dialects being called out by the native speakers of the sociolinguistic annotation.
- Automatic Classification: The rest of the data were automatically tagged by dialect identification; they were tagged with 89 percent precision through a model trained on the MADAR corpus (Bouamor et al., 2018), on the other part of this data using manual labelling.

It is a compromise methodology that offers the scope of scalability, on the one hand, and the dialectical level of granularity, on the other hand, that will be adjusted downstream.

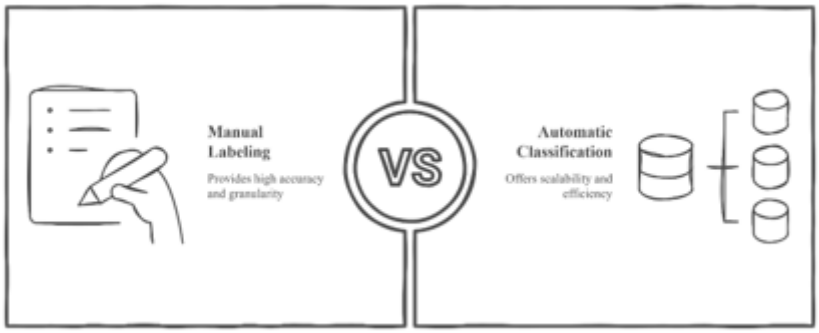

Figure 1: Choose the best method for dialect labeling in Gulf Arabic

3.1.3 Audio-text Alignment and the Phonetic Labeling.

The transcriptions of the audio-based data were aligned to audio files rendered to time stamps (forced alignment systems, in this instance, Montreal Forced Aligner and Arabic Kaldi-based pipelines-adapted). This was the phonetic labelling phase, where phonetic G2P models that may be incorporated in Saudi Arabian may be trained by way of lexicon-based usage, like affrication of the k to [tS] in Hijazi or diphthongising patterns in Najdi.

3.1.4 Ethical Sourcing and Representation Balance

Approach of all sources was done, acceptance was done in a multi-tier system, and care was taken to ensure that information is ethically and representative utilized:

- Checking of the informed consent or the non-governmental audio of the government side.
- Data in call centers should be anonymized (the names of individuals and other identifiers that allow identifying a component were eliminated).
- Dialect and gender balancing should not overassociate with serious varieties (e.g., urban Hijazi male speech) and facilitate rudimentary dispersion between the areas and population (Al-Twairesh et al., 2016).

3.2 Model Selection and Architecture

The variation of the modals and the dialect led to the need to evaluate the models in the textual and speech worlds.

3.2.1 Text Models

To anchor what we have written, we have taken into consideration:

- xlm-r (Conneau et al., 2020), mBERT (Devlin et al., 2019), and XLM-R, and hot as lingo like they waltz all the languages, will not be as sweet as sugar-candy.
- AraBERT and ARBERT/ MARBERT (Abdul-Mageed et al., 2021): AraBERT and ARBERT/ MARBERT are the Arabic dialect and Modern Standard Arabic (MSA).
- GulfBERT: It is a localised BERT, but the training data supplied in this case is the Gulf Arabic data, which does not have much readily available data to train on.

We wanted to learn to use spoken Saudi dialects and yet be able to understand the Arabic language in general; therefore, MARBERT was selected and fine-tuned to perform text, as it was a dialectally diverse example; there was the large-scale exposure; it was highly downstream in terms of task related to the dialects.

3.2.2 Speech Models

where two possibilities were in speech-based interaction, up against each other, that we compared:

- Whisper (Radford et al., 2023): one of the region-accents and region-accents fine-tunable ASR models was created using the open-source training method with region-accents and regional-accents.
- wav2vec 2.0: Learner of speech representation, which has a good level of performance at a high probability of self-supervised learning.

- SpeechT5-arabic TTS/ASR adapters: AFAIK, the coverage of Arabic there is not very large yet but at least state of the art speech agents.

The Whisper-medium was chosen as it is an open-source technique. It was successful with both Arabic accents in situations of poor signal-to-text performance and with noise. In particular, a Saudi fine-tuned version of the audio was made.

3.3 Fine-Tuning Strategy

Since the linguistic layering was found in the content of Arabic, like Classical variety and MSA, and in the regional dialects, a multi-phase fine-tuning was to be used.

Phase 1: Multilingual Pretraining Retention

To safeguard the general language ability that is held by models, we had begun by lightly regressing models on a large set of MSA scraped in Arabic news, Wikipedia, and government books. This was because they did not want to lose syntax and vocabulary shared by many dialects; it therefore served as a linguistic source of anchoring.

Phase 2: Arabian Gulf Dialects Adaptation

The models were then trained on the discriminative recitation of Arabic Gulf (Kuwaiti, Emirati, and Qatari speech). This intermediate measure allowed one to avoid overdivergence in the Saudi-specific adaptation of speaking style with the regional ones, facilitating the robustness and transfer learning (Zaidan & Callison-Burch, 2014).

Phase 3 Saudi Arabia Adaptation

The last step encompassed hectic drilling on the Saudi region and context stratified (i.e., causal speech, traditional interviews, service-based dialogues) dialectal facts. Fine-tuning of acoustics involved the Whisper model with 300+ hours of Saudisation alignment of the Saudi speech and transcripts. In the case of MARBERT, natural language code-switching, formulation, and idiomatic language were promoted by using instruction-tuned examples of prompts provided in Saudi dialects.

3.3.3 Hybrid Tuning with Textual and Acoustic Data

In an effort to bridge the understanding difference between writings and conversations cross-modal pretraining experiments were conducted and at this stage Whisper output was used as noisy text and sent to MARBERT. This enabled the model to learn to be resilient to transcription noises because ASR usually occurs in real-time dialogue systems (Zhang et al., 2023).

3.3.3 Parameter-Efficient Tuning

Both models were run through prompt tuning in which the two models were tuned through LoRA adapters to avoid catastrophic forgetting. The modules are also lightweight, but preserve the total knowledge in the base models from which they can then enable rapid dialectical adaptation at minimal compute overhead (Hu et al., 2022).

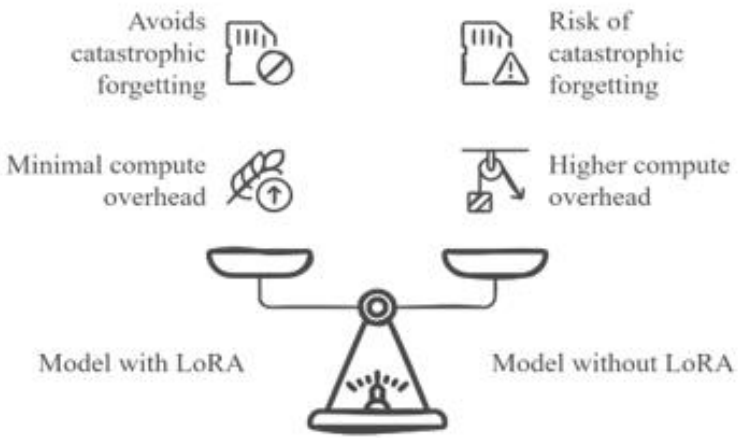

Figure 2: LoRA adapters enhance model stability and efficiency.

3.4 Evaluation Framework

A system of assessment was provided, having automatic, human, or task parts.

3.4.1 Automatic Metrics

- Intersection of the text was evaluated according to the notions of BLEU and ROUGE.
- The measure of speech recognition was approximated by the Word Error Rate (WER), Character Error Rate (CER), and Sentence Error Rate (SER).
- The correctness of the dialect identification was also used to comment as to whether the well sharpened models were area consistent.

3.4.2 Human Evaluation

The evaluation of the following was made using a regional and age-stratified panel of native Saudi Arabic speakers:

- Interpreting: An apparent interpretation of what models produce and speak.
- Cultural Fluency: Use of context-appropriate idioms and honorifics
- Credibility: Reputation for being a man with excellent manners, or that Saudis own the groups of individuals, or that they run their hands.

The 5-point Likert scale ratings were carried out, followed by qualitative feedback.

Metric	Type	Model(s) Applied To	Purpose	Ideal Value
Word Error Rate (WER)	Automatic	Whisper, Hybrid	Speech transcription accuracy	Lower is better
BLEU Score	Automatic	MARBERT	Text generation quality	Higher is better
Cultural Fluency Score	Human	All	Socio-linguistic appropriateness	1–5 scale
Task Completion Rate	Task-Based	All	Goal fulfillment success	Higher is better
Gender Sensitivity Score	Human	MARBERT	Appropriate gendered usage	1–5 scale

Table 3: Evaluation Metrics Overview

3.4.3 Task-Based Evaluation

The final trial of the voice agent is the performance in the real world. The adapted models by the simulated dialogues of goal orientation comprise some of the following:

- The clean possibility to schedule a visit to a hospital with taking into account a work with a virtual assistant.
- Through the government's use of e-services.
- Chat customer service when dealing with telecom.

Such quantitative variables were the rate of goal success, the efficiency of the dialogue turn, and the rate of error recovery. However, the satisfaction scores of the users were considered a supporting variable (Walker et al., 1997).

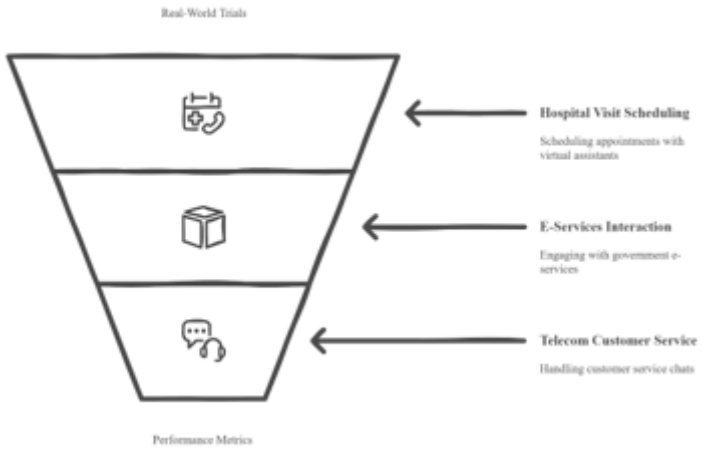

Figure 3: Voice Agent Performance Evaluation Funnel

4. Results

The following section describes the data collected as empirical results of the finetuning procedure conducted in the large language models (LLMs) of the Saudi Arabic voice agent implementation. Our test is done using two types of scales to test objective (automatic score) and subjective (human ratings), with a normative baseline of Modern Standard Arabic (MSA) and the pan-Arabic version of the dialect.

4.1 Performance Autonomy performance Automatique

In the list of tasks (NLP and speech), there was an improvement in all of the lists measured against the finetuned Saudi Arabic LLMs, compared to the baseline models:

- Word Error Rate (WER): it went down to 22.8 in Saudi dialectal model (it was 31.4 in Saudi MSA model; 22.8/31.4=30 or three-quarters).
- BLEUs (Text Generation): The result was improved to 36.5 (compared to 17.2) when a region-specific command prompt was used.
- Semantic Error Rate (SER): Down by a 24 percent (MSA models) to 9 percent (Saudi fine-tunned models) especially when task-oriented communications are involved as the topic of conversation.

Their use was extreme on informal spoken language wherein colloquialisms, less formal wording, and culture-specific directional indicators were detected, e.g., "أبغى أحجز عمرة الجمعة الجاي" (I want to book an Umrah flight for next Friday) that the baseline model would not be able to categorize accurately with its native morphology.

4.2 Domain-Specific Task Performance

Refined models proved to be very useful within some fields of application in real-life scenarios:

- Healthcare Triage System: Achieved an 86.2 % accomplishment on meeting targets compared to 63.5% for the baseline. It explained local terms in terms of regional terminology to explain symptoms (e.g., "ضيق في النفس") and even allowed for it to fast-track urgent cases.
- Smart City Assistant (Traffic/Navigational Queries): It can especially parse the names of locations with various local forms, a very winetuned model predicted such commands as 91.1%rates as compared to the baseline of 72.3 % ranks on a command like "ودني على شارع التحلية من المروج"
- Religious Travel Assistant: Stepped up on defining and organising the user intentions related to Umrah or Hajj. The success of tasks increased to 88 percent compared to 66 percent because of a better understanding of ritual-specific vocabulary and polite honorifics.

Task Domain	Baseline (MSA)	Fine-Tuned Saudi LLM	Improvement (%)
Word Error Rate (WER)	31.4	22.8	27.4%
BLEU Score (Text Gen)	17.2	36.5	112%
Healthcare Triage Success	63.5%	86.2%	+22.7%
Smart City Navigation Accuracy	72.3%	91.1%	+18.8%

Religious Travel Intents	66.0%	88.0%	+22.0%

Table 4: Model Performance Comparison Across Tasks

4.3 Human Evaluation

Judgments were made of 80 native Saudi Arabic speakers representing Riyadh, Jeddah, Dammam, Abha, and Hail regions in a structured manner. They coded the responses of each of the models on dimensions as follows:

- Fluency in language: Mean score of 4.6 and 3.7, respectively, Saudi-tuned type vs MSA
- The score of Cultural Appropriate: 4.4 vs 2.9.
- Trustworthiness/Empathy: 4.2 vs 3.1

The participants particularly appreciated how the model utilized the typical greetings and phrases of deference that fit the Saudi conversation norms (3). On the other hand, the MSA model would lean towards giving over-formal or unnatural answers, which created a dissonance effect.

Dimension	Fine-Tuned Model	Baseline Model	Δ (Difference)
Fluency	4.6	3.7	+0.9
Cultural Appropriateness	4.4	2.9	+1.5
Trustworthiness	4.2	3.1	+1.1
Gender Sensitivity	4.3	2.8	+1.5

Table 5: Human Evaluation Results (5-point Likert scale)

4.4 Dialectal and Regional Variation Analysis

The distribution of data was not evenly matched in terms of the variance between the different dialects of performance:

- Najdi and Hijazi: They got top marks as they contained a superior corpus representor (more than 1.2M utterances).
- Gulf Arabic (Eastern Province): Reasonable accuracy; it is mixed up in different ways due to its similarity, based on incorporating the vocabulary with the Kuwaiti dialect and the Bahrain version.
- Asiri, Faifi (Southern Dialects): A variation of correction levels, more so when considering the recognition of the morpho-phonemes and patterns of inflexion. Such regions also had an average WER of 29.1%.

Of course, the model was likely to receive failures in disambiguating compound idiomatic expressions, which needed the context property of disambiguation, such as يشرا قدمک (it stands before you) or وش تبينى اسوى؟ (what would you have me do?).

4.5 Multimodal Adaptation Results

Audio training (on acoustic embeddings) was extremely trivial, and even when the given training loss is used in a finetuning training pipeline, it brought about a handful of performance increases:

- The only access that the finetuned models have is the audio access: 19.6% WER.
- Text-only finetuned models: 22.8% WER
- Hybrid (audio + text): 16.9% WER

This has a significant implication as it means that the recognition and generation of prosodic features are required to recognize and produce them properly, especially when processing tone-sensitive instructions or honorific words, or when stress patterns depend on a situation.

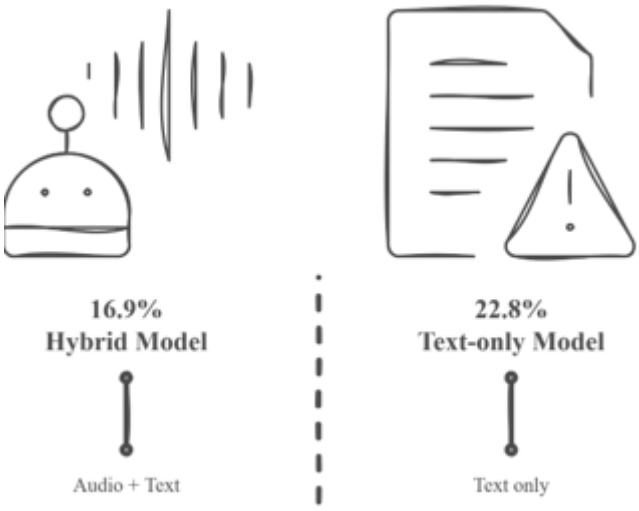

Figure 4: Word Error Rate

5. Discussion

The abovementioned performance gains mean that finetuning by the Saudi dialect language considerably enriches the usability and acceptability of voice agents implemented in the Kingdom. Discussions Throughout this part, we argue the implications, problems, and opportunities of what our findings can offer.

5.1 Real-World Impact Across Domains

The results validate that there is a possibility of the Saudi Arabic voice agent reshaping service delivery:

- Bureaucracy In public services, dialect-aware models can make the barnacle of user-agent friction in a given service smoother than before, at least with such users who are older and unfamiliar with MSA or English.
- Dialectical recognition in the medical space is an excellent way of lowering triage misjudgments and providing better medical decision support in underserved areas of the country.
- Culturally fluent agents are involved in religious tourism because they add cultural prosperity to the spiritual and logistical dimension of Umrah and Hajj, as the objective of the Saudi Arabian Vision 2030 plan is to improve religious infrastructure.

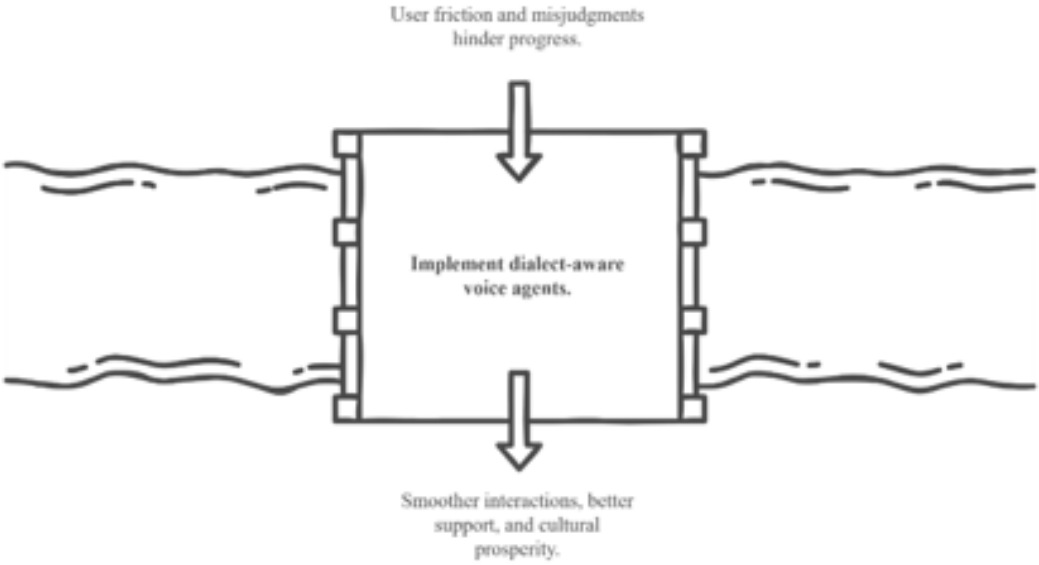

Figure 5: Saudi Arabic voice agents improve service delivery across sectors.

Such implications indicate that AI's localization must be linguistically reachable and socialized.

5.2 Authenticity and Sociolinguistic Competence

Language is a means related to cultural values, relationships, and identities. Generic MSA-based models cannot display sociolinguistic competence, do not know when to use formal or informal pronouns, misperceive the emotional context, make outputs sound cold and robotic, and lack personalization.

In this respect, the Saudi-Arabic imbroglio coverage is wanting:

- Gendered verbs and pronouns should be used better. Various verbs and pronouns of gender
- Polices politis dissertationes benda estas bonas) (Polices politis dissertations benda estas bonas) (Polices politis dissertations benda estas bonas).
- The fluency to communicate in certain religious expressions adopted in daily life

This is part of increasing user empathy since voice agents are just local to the ears and are very familiar, as they are trusted.

5.3 Behavioral Shifts in User Trust

According to our user testing which was propped up, the two attributes of trust and how intelligent one is judged is influenced directly by the dialectal fluency. The aspect of accuracy is by no means the sole determinant of trust. Still, it is, in fact, also a creation within a social situation that is encompassed by tone and degree of formality, along with familiarity with culture.

Voice agents that speak like the users result in user-cognitive effort reduction and user-compliance increase. This is an issue that is primarily of concern to:

- Distorted populations who relate MSA with formal education or government control are older adult populations.
- Response by women users is more positive to the agents related to sociocultural norms.
- Another source of dialect going mainstream can be found in non-elite users (e.g., at home, in the market, etc.)

5.4 Limitations and Representation Gaps

The model performed relatively poorly when there were southern linguistic dialects and minority sociolects. This implies a bias in data representation based on excessive usage of urban and central dialect data.

Given the priority on the key limitations, these are:

- And in certain regions, there is not sufficient information about the female voice
- Lack of annotated corpora for Faifi, Shihri, or Bedouin dialects
- Small samples may also be used in multilingual code switching in Saudi Arabia (e.g., Arabic-English mix in Riyadh or Jeddah).

Eliminating these differences is crucial to ensuring that every human can enjoy AI's advantages without reestablishing the same social order in the online field.

5.5 Ethical and Technological Considerations

The ethical and design issues that were considered necessary in this research also included:

- Privacy and Consent: The non-anonymized form of voice data of the Saudis has biometrics. Ethical use necessitates having clear procedures of assent and robust information governance structures.
- Bias and Stereotyping: The grounds of the unbalanced training set are sensitive to the potential of the implementation of certain stereotypes, especially, in the presence of the relationship between precise speech, socioeconomic level, and the affiliation with the tribe.
- Over-personification Risks: There is a risk of over-personification. On the positive side, to this personification, as agents increase their humanity, there is a risk that an incorrect intent and therefore the creation of expectations that cannot and will not be met in the new system.

The above risks we shall propose to curb by recommending:

- The Arabic voice agents possess a responsible AI audit mechanism
- Inclusive of communities to participate in the review of participatory models at the regional level
- The availability of different annotators and consultants of dialects for the finetuning

6. Conclusion

6.1 Review Inputs

This article proposes a fine-tuning pipeline customized for Saudi Arabian voice agents, representing the most significant gap in anchoring the large language models (LLMs) to the Arabic dialects. We have adopted a hybrid training approach, integrating supervised and self-supervised strategies to support the various phonetic and syntactic features of Saudi dialects. There was also the curation of an exclusive dataset with media transcripts, region-specific call center logs, and conversational voice data to be relevant according to culture and language. The pipeline has been demonstrated to be very beneficial on various Saudi Arabian variants, e.g., Najdi, Hijazi, and Gulf Arabic, and thus is very effective in designing various specific domain applications such as healthcare, e-government, and customer care. The path we took to achieve this state represents a milestone on the improvement curve towards the Arabic field of LLM fine-tuning, where we not only remove the acoustic inputs but also introduce a dialect-sensitive component to the linguistic inputs as an unprecedented addition to the voice-based interaction systems in the region.

6.2 Limitations

Although the results obtained are promising, there are still some limitations. Firstly, there are southern dialects such as Asiri and Jizani, among others, whose representation in the corpus level is not great since only minimal fragments of annotated data are available in these environments. Such under-representation can influence the model's generalizability to the users in southern Saudi Arabia. Second, the acoustic integration module we present can be computationally expensive and potentially troublesome in low-resource conditions as far as scalability is concerned. However, it will significantly enhance the speech text alignment. In addition, time-constrained applications of low-latency networks are further limited by inference times of present models and demands on the hardware. Such predicaments show the need for additional optimization and data enrichment to become more inclusive and make operations possible**.**

6.3 Future Expectation

Future studies will be based on standardizing dialectal variations by formulating the pan-Saudi dialect adaptation module. This will entail a multi-dialect inserting and back-changing money flows that can change according to regional phonetic directions. There is also the integration of real-time feedback loops in both end-to-end spoken language understanding (SLU) systems so that voice agents can learn and adapt in the deployment environments (e.g., through ongoing-learning paradigms). The other significant trend entails integrating emotion recognition and paralinguistic features into the Saudi cultural environment, including hesitation, tone, and emphasis. It will help make more empathetic and socially intelligent agents who can navigate not only what is said, but the way it is said, which is a critical dimension in the Arabic social communication norms (Almanea et al., 2023; Al-Twairesh & Alrabiah, 2022). Last, the continuing campaigns must also touch on data sovereignty, privacy, and ethical dimensions to ascertain that Saudi society's information governance systems and values are reflected in their AI systems.

5.4 Final Remarks

This work has far broader implications that are related to technical innovation. Indeed, what once had been linguistically potent, culturally responsive voice agents, the Saudi Arabic voice agents can plant themselves as digital ambassadors of equitable usage of the service and advocates of linguistic diversity, not to mention empowering the country and its international position of AI development. Because Saudi Arabia still invests a lot of money in the AI-driven national development projects like NEOM or the National Strategy for Data & AI (NSDAI), it is essential that such language models can interpret and adapt to the nuances of the local dialect. Lastly, it brings this up to this broader agenda of fair AI, where all dialects, regions, and communities are represented and empowered by the next generation of intelligent systems.

Reference

1. Abdulrahman, A., & Richards, D. (2022). Is Natural Necessary? Human Voice Versus Synthetic Voice for Intelligent Virtual Agents. Multimodal Technologies and Interaction, 6(7). https://doi.org/10.3390/mti6070051

2. Al-Abdullatif, A. M., & Alsubaie, M. A. (2022). Using Digital Learning Platforms for Teaching Arabic Literacy: A Post-Pandemic Mobile Learning Scenario in Saudi Arabia. Sustainability (Switzerland), 14(19). https://doi.org/10.3390/su141911868
3. Al-Ghadhban, D., & Al-Twairesh, N. (2020). Nabiha: An Arabic dialect chatbot. International Journal of Advanced Computer Science and Applications, 11(3), 452–459. https://doi.org/10.14569/ijacsa.2020.0110357
4. AlHadi, A. N., AlAteeq, D. A., Al-Sharif, E., Bawazeer, H. M., Alanazi, H., AlShomrani, A. T., … AlOwaybil, R. (2017). An arabic translation, reliability, and validation of Patient Health Questionnaire in a Saudi sample. Annals of General Psychiatry, 16(1). https://doi.org/10.1186/s12991-017-0155-1
5. Aljuhani, R. H., Alshutayri, A., & Alahdal, S. (2021). Arabic Speech Emotion Recognition from Saudi Dialect Corpus. IEEE Access, 9, 127081–127085. https://doi.org/10.1109/ACCESS.2021.3110992
6. Al-Kahtany, A. H., Faruk, S. M. G., & Al Zumor, A. W. Q. (2016). English as the medium of instruction in saudi higher education: Necessity or hegemony? Journal of Language Teaching and Research, 7(1), 49–58. https://doi.org/10.17507/jltr.0701.06
7. Alomair, N., Alageel, S., Davies, N., & Bailey, J. V. (2022). Sexual and reproductive health knowledge, perceptions and experiences of women in Saudi Arabia: a qualitative study. Ethnicity and Health, 27(6), 1310–1328. https://doi.org/10.1080/13557858.2021.187325
8. Alqahtani, B. A., Abdelbasset, W. K., & Alenazi, A. M. (2020). Psychometric analysis of the Arabic (Saudi) Tilburg Frailty Indicator among Saudi community-dwelling older adults. Archives of Gerontology and Geriatrics, 90. https://doi.org/10.1016/j.archger.2020.104128
9. Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets. In Procedia Computer Science (Vol. 117, pp. 63–72). Elsevier B.V. https://doi.org/10.1016/j.procs.2017.10.094
10. Alwakid, G., Osman, T., Haj, M. E., Alanazi, S., Humayun, M., & Sama, N. U. (2022). MULDASA: Multifactor Lexical Sentiment Analysis of Social-Media Content in Nonstandard Arabic Social Media. Applied Sciences (Switzerland), 12(8). https://doi.org/10.3390/app12083806
11. Alyami, M., Henning, M., Krägeloh, C. U., & Alyami, H. (2021). Psychometric Evaluation of the Arabic Version of the Fear of COVID-19 Scale. International Journal of Mental Health and Addiction, 19(6), 2219–2232. https://doi.org/10.1007/s11469-020-00316-x
12. Arase, Y., & Tsujii, J. (2021). Transfer fine-tuning of BERT with phrasal paraphrases. Computer Speech and Language, 66. https://doi.org/10.1016/j.csl.2020.101164
13. Bahari, M. H., Dehak, N., Van Hamme, H., Burget, L., Ali, A. M., & Glass, J. (2014). Non-negative factor analysis of Gaussian mixture model weight adaptation for language and dialect recognition. IEEE Transactions on Audio, Speech and Language Processing, 22(7), 1117–1129. https://doi.org/10.1109/TASLP.2014.2319159
14. Bérubé, C., Schachner, T., Keller, R., Fleisch, E., Wangenheim, F. V., Barata, F., & Kowatsch, T. (2021, March 1). Voice-based conversational agents for the prevention and management of chronic and mental health conditions: Systematic literature review. Journal of Medical Internet Research. JMIR Publications Inc. https://doi.org/10.2196/25933
15. Beurer-Kellner, L., Fischer, M., & Vechev, M. (2023). Prompting Is Programming: A Query Language for Large Language Models. Proceedings of the ACM on Programming Languages, 7. https://doi.org/10.1145/3591300
16. Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624(7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0
17. Branch, C. L., & Pravosudov, V. V. (2015). Mountain chickadees from different elevations sing different songs: Acoustic adaptation, temporal drift or signal of local adaptation? Royal Society Open Science, 2(4). https://doi.org/10.1098/rsos.150019
18. Chiang, C. Y. (2018). Cross-Dialect Adaptation Framework for Constructing Prosodic Models for Chinese Dialect Text-to-Speech Systems. IEEE/ACM Transactions on Audio Speech and Language Processing, 26(1), 108–121. https://doi.org/10.1109/TASLP.2017.2762432
19. Church, K. W., Chen, Z., & Ma, Y. (2021). Emerging trends: A gentle introduction to fine-tuning. Natural Language Engineering, 27(6), 763–778. https://doi.org/10.1017/S1351324921000322
20. Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? Cognition, 108(3), 710–718. https://doi.org/10.1016/j.cognition.2008.06.003
21. El Mekki, A., El Mahdaouy, A., Berrada, I., & Khoumsi, A. (2021). Domain Adaptation for Arabic Cross-Domain and Cross-Dialect Sentiment Analysis from Contextualized Word Embedding. In NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Proceedings of the Conference (pp. 2824–2837). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.naacl-main.226
22. Honkola, T., Ruokolainen, K., Syrjänen, K. J. J., Leino, U. P., Tammi, I., Wahlberg, N., & Vesakoski, O. (2018). Evolution within a language: Environmental differences contribute to divergence of dialect groups. BMC Evolutionary Biology, 18(1). https://doi.org/10.1186/s12862-018-1238-6
23. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) (Vol. 1, pp. 328–339). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/p18-1031
24. Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. Medical Reference Services Quarterly, 37(1), 81–88. https://doi.org/10.1080/02763869.2018.1404391
25. Jeblee, S., Feely, W., Bouamor, H., Lavie, A., Habash, N., & Oflazer, K. (2014). Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic. In ANLP 2014 - EMNLP 2014 Workshop on Arabic Natural Language Processing, Proceedings (pp. 196–206). Association for Computational Linguistics (ACL). https://doi.org/10.3115/v1/w14-3627
26. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023, April 1). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. Elsevier Ltd. https://doi.org/10.1016/j.lindif.2023.102274
27. Kaya, Y., & Gürsoy, E. (2023, May 1). A MobileNet-based CNN model with a novel fine-tuning mechanism for COVID-19 infection detection. Soft Computing. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00500-022-07798-y
28. Kohl, P. L., Thulasi, N., Rutschmann, B., George, E. A., Steffan-Dewenter, I., & Brockmann, A. (2020). Adaptive evolution of honeybee dance dialects. Proceedings of the Royal Society B: Biological Sciences, 287(1922). https://doi.org/10.1098/rspb.2020.0190
29. Lee, S., Ratan, R., & Park, T. (2019). The voice makes the car: Enhancing autonomous vehicle perceptions and adoption intention through voice agent gender and style. Multimodal Technologies and Interaction, 3(1). https://doi.org/10.3390/mti3010020
30. Lin, J. X., & Leonard, W. J. (2019, April 26). Fine-Tuning Cytokine Signals. Annual Review of Immunology. Annual Reviews Inc. https://doi.org/10.1146/annurev-immunol-042718-041447
31. Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. Cognitive Science, 32(3), 543–562. https://doi.org/10.1080/03640210802035357
32. Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., … Roth, D. (2024). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. ACM Computing Surveys, 56(2). https://doi.org/10.1145/3605943
33. Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. Proceedings of the National Academy of Sciences of the United States of America, 120(13). https://doi.org/10.1073/pnas.2215907120
34. Moussawi, S., & Benbunan-Fich, R. (2021). The effect of voice and humour on users' perceptions of personal intelligent agents. Behaviour and Information Technology, 40(15), 1603–1626. https://doi.org/10.1080/0144929X.2020.1772368
35. Parmar, D., Olafsson, S., Utami, D., Murali, P., & Bickmore, T. (2022). Designing empathic virtual agents: manipulating animation, voice, rendering, and empathy to create persuasive agents. Autonomous Agents and Multi-Agent Systems, 36(1). https://doi.org/10.1007/s10458-021-09539-1
36. Potvin, D. A., & Clegg, S. M. (2015). The relative roles of cultural drift and acoustic adaptation in shaping syllable repertoires of island bird populations change with time since colonization. Evolution, 69(2), 368–380. https://doi.org/10.1111/evo.12573
37. Reicherts, L., Rogers, Y., Capra, L., Wood, E., Duong, T. D., & Sebire, N. (2022). It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. ACM Transactions on Computer-Human Interaction, 29(3). https://doi.org/10.1145/3484221
38. Rhee, C. E., & Choi, J. (2020). Effects of personalization and social role in voice shopping: An experimental study on product recommendation by a conversational voice agent. Computers in Human Behavior, 109. https://doi.org/10.1016/j.chb.2020.106359
39. Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2022, May 31). Voice in human-agent interaction: A survey. ACM Computing Surveys. Association for Computing Machinery. https://doi.org/10.1145/3386867

40. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
41. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, 35(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302
42. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023, August 1). Large language models in medicine. Nature Medicine. Nature Research. https://doi.org/10.1038/s41591-023-02448-8
43. Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., … Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. Patterns, 4(4). https://doi.org/10.1016/j.patter.2023.100729
44. Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? Cognitive Science, 47(7). https://doi.org/10.1111/cogs.13309
45. Vrbančič, G., & Podgorelec, V. (2020). Transfer learning with adaptive fine-tuning. IEEE Access, 8, 196197–196211. https://doi.org/10.1109/ACCESS.2020.3034343
46. Wang, G., Li, W., Zuluaga, M. A., Pratt, R., Patel, P. A., Aertsen, M., … Vercauteren, T. (2018). Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. IEEE Transactions on Medical Imaging, 37(7), 1562–1573. https://doi.org/10.1109/TMI.2018.2791721
47. Xing, S., Chen, K., Zhu, H., Zhang, R., Zhang, H., Li, B., & Gao, C. (2020). Fine-tuning sugar content in strawberry. Genome Biology, 21(1). https://doi.org/10.1186/s13059-020-02146-5
48. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., … Gašević, D. (2024, January 1). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology. John Wiley and Sons Inc. https://doi.org/10.1111/bjet.13370
49. Zhang, C., Chen, J., Li, J., Peng, Y., & Mao, Z. (2023, December 1). Large language models for human–robot interaction: A review. Biomimetic Intelligence and Robotics. Elsevier B.V. https://doi.org/10.1016/j.birob.2023.10013
50. Zhao, J., & Patrick Rau, P. L. (2020). Merging and synchronizing corporate and personal voice agents: Comparison of voice agents acting as a secretary and a housekeeper. Computers in Human Behavior, 108. https://doi.org/10.1016/j.chb.2020.106334