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

Reducing Hallucination in Multilingual Voice Agents Using Instruction-Tuned Models

Corresponding Author: Mahmoud Abdelhadi Mahmoud Safia

Corresponding Email: msafia@jccs.com.sa

Bachelor's degree in computer engineering, Aleppo University

Abstract

In highly applied multilingual voice agents of customer service and interactive AI systems in the world, one persistent problem constantly haunts the industry/field: hallucinations--syntactically adequate responses, which are logically wrong or simply inapplicable. These are multiplied in the multilingual setting as they have disparate linguistic peculiarities, professional vocabularies, and underrepresented languages. The current paper investigates how far it is forthcoming to trim hallucinations in multilingual voice agents via instruction-tuned models finetuned to pay heed to clear task guidelines. We provide finetuning of the big-resource multilingual transformer-based models (that is, mT5, XGLM, BLOOMZ), and propose the measure across ten languages as the measure for understanding, factual accuracy, and language consistency.

To approximate the uniqueness of our methodology, we propose a hybrid evaluation apparatus of automated rating (BLEU, COMET, and factual consistency scores) and human evaluation, which is modified according to the cultural and linguistic peculiarities of the language, which is the subject of our research. Experiments conducted by us indicate that the tuning of instruction by a large margin decreases hallucinations, particularly when referring to retrieval-augmented generation (RAG) and task-completion instructions. We also discuss how the phenomenon of instruction tuning used in high-resource and low-resource languages is different and may lead to hallucination in the context of various families of languages. It may also be performed to determine the trends on both the structural and syntactic levels. Last, we suggest the most effective way to calibrate the pipes with the instructions in multilingual voice systems on an operational stage.

With our findings, the future success of multilingual voice assistance is adapting instruction-tuned models to generate more factual reliability, compatibility in the languages, and confidence being guaranteed through mediation of safer and more responsible conversational AI agents across the language divide.

Keywords- multilingual voice agents, instruction-tuning, hallucination reduction, factual consistency, conversational AI

1. Introduction

The LLMs transformed natural language processing by providing an efficient dialogue system and voice agents. However, it does have one deadly sin that has not been addressed yet: hallucination or the generation of plausible contents, except that they are inaccurate or false (Ji et al., 2023). In voice agents, the hallucinations may undermine the user's trust in the product, cause misinformation, and potentially severe safety issues, particularly in healthcare, finances, or customer service. The problem of hallucinations is further aggravated when multilingualism is involved because language ambiguity, the disparity in translation, and the lack of quality training data to represent low-resource languages all contribute to it (Ruder et al., 2021; Costa-jussà et al., 2022). The voice agents have to think or produce content in multiple languages with different types of syntax, cultural conventions, and idiomatic expressions that do not always align well with mostly English-focused pretraining data of most LLMs. Therefore, hallucination rates in non-English interactions are more likely to appear in morphologically rich or typologically distant languages (Anastasopoulos & Neubig, 2020).

It is necessary to decrease hallucinations to achieve factual reliability, the safety of users, and the worldwide scalability of voice agents. Where access to information occurs in multilingual contexts, e.g., cross-border customer care, voice-activated assistants speaking the local dialect, the user should feel confident that they can glean accurate information, not just fluent information.

This research identifies whether the use of instruction-tuned models presents a possible solution to minimize the occurrence of hallucinations in multilingual voice agents. The training of LLMs with instructions is a method related to instruction

tuning: It involves fine-tuning on the datasets, collecting task-specific prompts and grounded responses so that LLMs are better able to adhere to user intent and resist producing groundless information (Wei et al., 2022; Chung et al., 2022). Instruction tuning has improved alignment and task generalization in monolingual settings. Still, the effect on the prevalence of the hallucination phenomenon in multilingual voice-based dialogue systems has not been studied.

1.1 Research Objective

The study will systematically review how instruction-tuned models reduce hallucinations in various languages regarding voice application in agents. In particular, the following questions are answered:

- What does the rate of hallucinations depend on in different languages, as well as tasks, in standard versus instruction-tuned models?
- Which are the best instruction tuning strategies that can minimize multilingual hallucinations?
- Will an evaluation scheme using both automatic and human measures be able to measure the severity of hallucination in multilingual settings?

1.2 Contributions

To contribute to the existing knowledge in this area, we offer the following contributions:

1. Comparing Hallucination Rates: We compare the frequency and severity of hallucinations in a heterogeneous language spectrum (with low-resource languages) on both artificial and real-world multilingual voice agent processing tasks.
2. Instruction tuning strategies: We investigate some multilingual instruction tuning strategies, zero-shot, few-shot, and culturally adaptive prompting, and ascertain their consequences on factuality.
3. Evaluation Framework: We present a multilingual (pre) training based hallucination evaluation pipeline that integrates automated (e.g. BERTScore, factual consistency) with human metrics rated by experts on the fluency, factuality and trustworthiness scales.

This research opens up the potential to create more compelling, linguistically diverse voice agents by reconciling the two domains of instruction-guided modeling methods and multilingual dialogue assessment. It provides a roadmap to subsequent deployment of language model systems with low hallucination in low-resource settings.

2. Background and Related Work

2.1 Hallucination in Neural Language Models

In neural language models, hallucination is seen to mean the production of text which is fluent and yet factually false or uncorrupted. It is especially an issue whenever the stakes are high, as is the case in healthcare, legal, and multilingual customer services, where fabrication may bring grave ramifications.

There can be two large types of hallucinations: intrinsic and extrinsic (Ji et al., 2023). Intrinsic hallucinations occur when the output reveals some facts that are not explicitly contained in the input, such as when a summarization model invents one not stated by the source. On the other hand, extrinsic hallucinations create possible but unverifiable information, likely because of the lack of grounding or because it was overgeneralized.

Hallucination has several causes. These generalizations may be the falseness of the model encoded because data bias may give a skewed view of the world, such as the over-representation of English-centric or Western views (Min et al., 2022). Underresourced or any especially morphologically rich language leads to a greater likelihood of misinterpreting or overgenerating due to language ambiguity. Moreover, the neural models are overconfident in the low-probability forecasts and can strengthen the hallucinations during decoding (Kadavath et al., 2022).

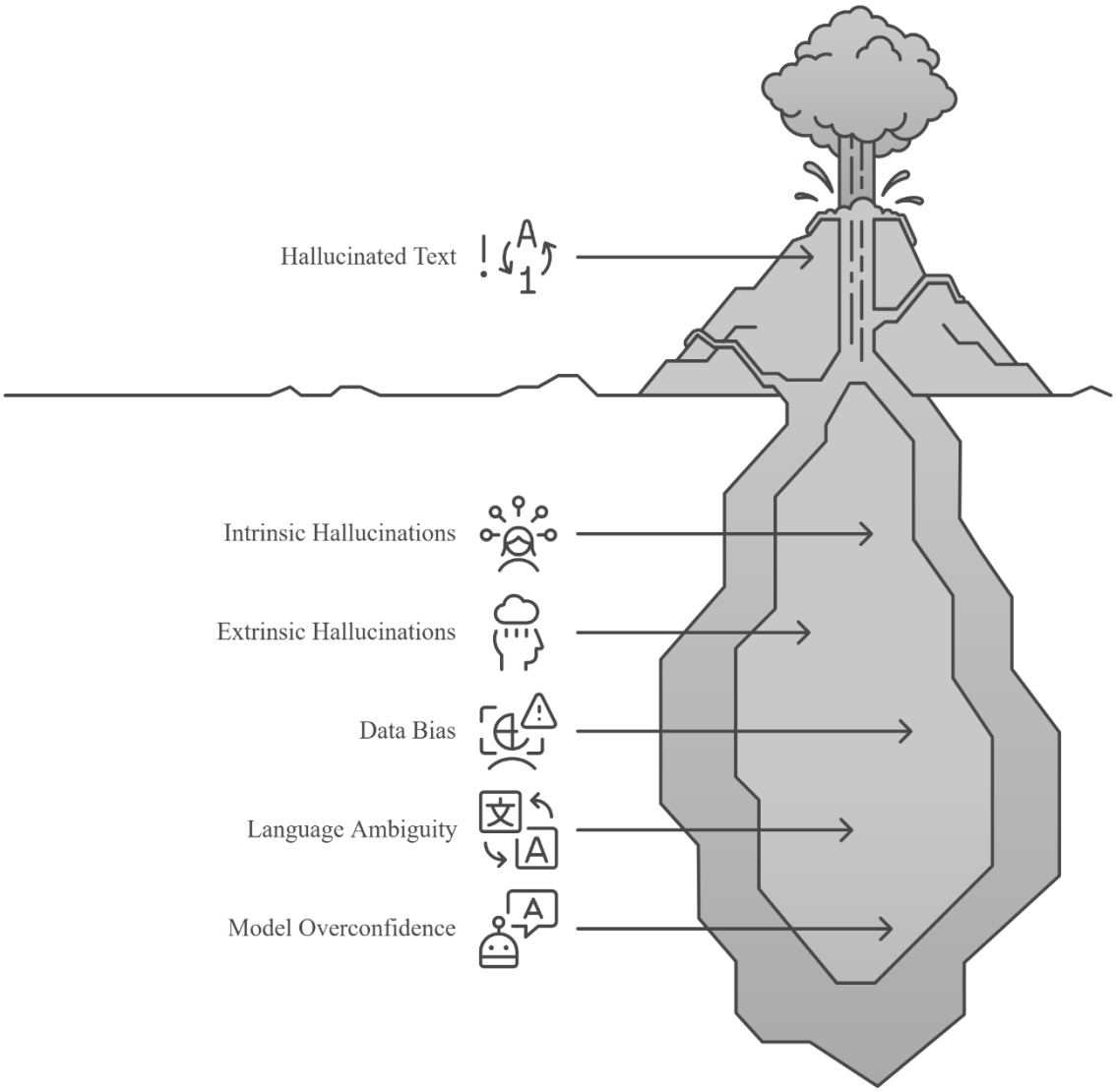

Figure 1: Hallucinations in Neural Language Models: Unveiling the Hidden Causes

2.2 Multilingual Voice Agents

Multilingual voice agents comprise automatic speech recognition (ASR), natural language understanding (NLU), and text-to-speech (TTS) interventions to aid in cross-support. Such systems have extensive applications in call centers (customer service), health (medical triage), education, and personal assistants. Nevertheless, voice agents hinder hallucination at this point because of the increased danger of these scenarios caused by real-time communication and a perceived font of authority of the voice (Zhao et al., 2023).

The issue of multilingual NLP has peculiarities. The difference in performance across the languages is well known. It exists because of inefficiencies in tokenization (e.g., the subword fragmentation of agglutinative languages, such as Turkish), morphosyntactic complexity (e.g., case marking in Slavic languages), and the insufficiently annotated corpora of most low-resource languages (Conneau et al., 2020). Such problems increase the prevalence of hallucinations in a multilingual environment and, most importantly, one that involves unbalanced training data. Moreover, any cultural or linguistic misplacement, e.g., by trying to translate idioms, ignoring pragmatic indicators, etc., can cause a hallucinated reaction, whose content will be semantically nonsensical or even offensive in a particular situation (Ponti et al., 2021).

2.3 Tuned- instructions Models

Tuning is an effort to bring language models closer to the intention of human beings by fine-tuning the instructions posed in a task-specific manner and presented in a natural form. Brilliant at zero-shot and few-shot performance on many NLP tasks was T5 (Raffel et al., 2020), FLAN-T5 (Chung et al., 2022), mT5 (Xue et al., 2021), and BLOOMZ (Muennighoff et al., 2023).

The instruction-tuned models are especially encouraging in multilingual ones as they incorporate their capacity to learn to follow organized prompts and conform more to the factual feedback input. Respectively, FLAN-T5 is more factual than T5 due to training to condition on the instructions format (Chung et al., 2022). Equally, BLOOMZ has been trained on cross-linguistic prompts in 46 languages and indicates potential to perform controlled generation.

Nevertheless, the existing multilingual instruction-tuning initiatives are relatively small-scale and ineffective. The English-dominant or machine-translated without paying much attention to cultural or syntactic accuracy, there are many instruction-tuning datasets (Zhou et al., 2023). Cross-linguistic transfer learning is not uniform, and results suffer significantly in low-resource languages.

Furthermore, immediate compliance and empirical basing are not ensured. Although models obey instructions, they can hallucinate when confronted with puzzling invocations or when trained on conflict-pairs of instructions and tasks (Honovich et al., 2022).

2.4 Evaluation of Hallucination

The granulation of hallucination is a research challenge that has yet to be solved in the multilingual setting. Several benchmarks are suggested:

- TruthfulQA (Lin et al., 2022): Trial of the knowledge based on the confidence in the veracity of the response to the generally accepted notions in the English language and people's erroneous beliefs.
- The factuality Questions: How many times a claim can be deemed as being factually consistent, checking it for incorrectness using factuality into metrics: Automatic, generally binary counts (e.g., BERTScore, FactCC, QuestEval), which may be used to determine factual consistency of the claim.
- QAGS and SummaC: They were created to perform the task of summarization but have since been applied more on hallucination detection.

There is an immediate need for culturally adaptive assessment procedures for language understanding. It consists of multilingual factuality benchmarks, crowd-sourced/expert-annotated datasets, and heterogeneous human/AI test pipelines capable of recognizing sociolinguistic and pragmatic variation in underrepresented languages.

Although successful in a monolingual context, these benchmarks have critical drawbacks when articulated to a multilingual voice agent. One is that most of them depend on English-oriented corpora or inferences concerning logical entailment, which are not cross-linguistically applicable to lexicon differences in meaning and syntax. Second, cultural background and language specifics are disregarded, e.g., a remark found wrong in one language might have entailed truth in another locally due to local information or a locational difference in meaning (Asai et al., 2021).

3. Methodology

The study will undertake a multi-step methodology that addresses the problem of reducing the problem of hallucination in multilingual voice agents deployed through language models learned through instructions. The four major components of this method will entail the following: (1) language and dataset choice, (2) instruction tuning pipeline, (3) evaluation, and (4) comparison with a baseline system.

3.1 Language/Dataset Selection

To determine whether our decisions were as generalizable as possible, we selected ten languages (English, Spanish, Arabic, Hindi, Mandarin, Russian, Swahili, Vietnamese, Turkish, and Finnish) to reflect as diverse a language family as possible, the setting in which they are used, and access to materials. This sample represents the Indo-European, Afro-Asiatic, Turkic, Sino-Tibetan, and Niger-Congo families and, thus, allows evaluation of hallucination in typologically and morphologically diverse systems (Ponti et al., 2020).

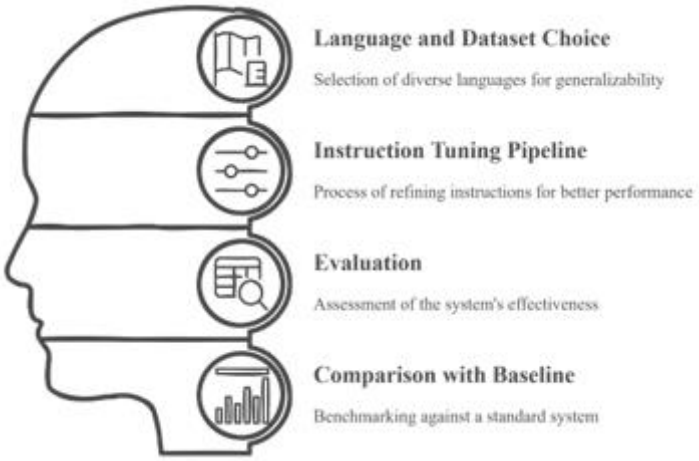

Figure 2: Methodology for Reducing Hallucination in Voice Agents

The selection of the datasets prioritized the idea of diversity of tasks and factual anchoring, where structured dialogs and knowledge-intensive tasks were added to them. These corpora participating are:

- Task-oriented dialogue corpora: MultiWOZ 2.4 (Zang et al., 2020) contained multilingual goal-oriented dialogues across several domains (travel, weather, etc.). MASSIVE (FitzGerald et al., 2022) provided multilingual goal-oriented dialogues across various domains (travel, weather, etc.).
- Factual QA datasets: Natural Questions, TyDi QA (Clark et al., 2020), and XQuAD, were used to verify factual consistency and cross-lingual question answering.
- Translation and paraphrase datasets: Tatoeba and PAWS-X (Yang et al., 2019) assisted in cross-lingual alignment and paraphrase comprehension that are important to evaluate hallucination in rephrasings.

State-of-the-art models translated all datasets as required and normalized them, where manual validation of a stratified subset was performed, so as not to lose fidelity.

Language	Language Family	Resource Level	Scripts Used	NLP Challenges	Dataset(s)	Dialogue Domain(s)
English	Indo-European	High	Latin	Verb tense ambiguity	MultiWOZ, Natural Questions	Travel, Hotels, QA
Spanish	Indo-European	High	Latin	Gender agreement, morphology	XQuAD, MASSIVE	Weather, Booking
Arabic	Afro-Asiatic	Medium	Arabic	Root-based morphology, dialect variation	TyDi QA, PAWS-X	Government, QA
Hindi	Indo-European	Medium	Devanagari	Free word order, compound verbs	TyDi QA	Tourism, Health
Mandarin	Sino-Tibetan	High	Simplified Chinese	Tonality, word segmentation	Tatoeba, TyDi QA	Education, QA

Russian	Indo-European	High	Cyrillic	Case inflection, aspectual verbs	PAWS-X, MASSIVE	Banking, Summary
Swahili	Niger-Congo	Low	Latin	Agglutination, noun classes	TyDi QA, MASSIVE	Weather, QA
Vietnamese	Austroasiatic	Low	Latin (with diacritics)	Tonality, isolating morphology	Tatoeba, MASSIVE	Local services
Turkish	Turkic	Medium	Latin	Agglutinative structure, vowel harmony	PAWS-X, MASSIVE	Transport
Finnish	Uralic	Low	Latin	Long compound words, inflections	MASSIVE, XQuAD	Healthcare, QA

Table 1: Language Characteristics and Dataset Overview

3.2 Instruction Tuning Pipeline

The primary mechanism we used to minimize hallucination is instruction. This was done by aligning already prepared multilingual models to be more faithful in executing the tasks and languages based on natural language instructions.

3.2.1 Model Selection

Three experiments with multilingual models were chosen as being instruction-tunable:

- mT5 (Xue et al., 2021): mC4-adapted T5 that can work in the text-to-text application as it is a multilingual model.
- XGLM (Open research, 2022): Optimized in terms of generative model, low-resource, and cross-lingual friendly.
- BLOOMZ (Muennighoff et al., 2023): optimized according to the compounded prompting of a cross-lingual model whose training has been based on the instruction- refinement- tuning paradigm of how BLOOM is instructed.

These models represent the spectrum of pretraining strategies, magnitude, domains and allow us to gain insights into the connection between architectures, tuning and hallucination.

3.2.2 Preprocessing and Alignment

All the training data was standardized into the same format of instruction-response. For example:

Would you use your own words and say it in your way--to your way of saying these words, in Arabic: The cat sat on the mat.

Response:" قط جلس على السجادة."

Cultural and syntactic naturalness of instructions translation and localization information was carried out by native speakers. In cases where they were possible, one used task-type metadata to facilitate context-aware tuning (e.g., preceding the task name, e.g., "Question Answering:") to the extent practicable.

3.2.3 Prompting Techniques

To make maximum generalization, we applied three prompting strategies both in training and evaluation:

- Zero-shot prompting: Instead of examples, there are instructions.
- Few-shot prompting - 24 instantiated examples of an action.
- Task prefixing: Some identifiers of the particular task (e.g., in their example, the words that precede the heads of prompts, such as Translate or Summarize, are placed in advance of the heads of prompts (Wei et al., 2022).

Such methods aid in unraveling hallucination due to inadequate grounding of the hallucinating false perception, and the shortcomings in the generalization of tasks.

3.2.4 Fine-Tuning Strategy

Instructions on the models were given using the transformers package in Hugging Face and performing model-specific tuning. Hyperparameters were kept constant across models to provide some form of ground for comparison:

- Learning rate:3 -5
- Batch size: 64 (Gradient accumulation turned on because of savings of memory)
- Epochs: 5 (as well as early stopping on validation on a factual basis)
- Max size for input is 512. Max size for output: 128 tools.
- Optimiser: AdamW, weight decay: 0.01

It was pretrained on combining 4 A100 GPUs in mixed-precision (fp16). The experiments were replicated because there were limited random seeds and uniform tokenization.

Model	Parameters	Pretraining Corpus	Instruction Source	Precision
mT5	580M	mC4 multilingual	Self-generated + XINSTRUCT	FP16
XGLM	564M	CC100	Prompt-translated + English	FP16
BLOOMZ	1.1B	ROOTS	Cross-lingual prompt pairs	FP16

Table 2: Instruction Tuning Configurations by Model

3.3 Evaluation Design

In an attempt to be rigorous in quantifying hallucination reduction, we used a combination of automatic and human metrics, both on the surface quality and deep factual alignment.

3.3.1 Automatic Metrics

We have employed complementary pairs of measurements as an evaluation of the quality of generation:

- BLEU/ROUGE:Reflink: portion of reference of words word to word (Papineni et al., 2002; Lin, 2004).
- chrF++/COMET: similarity and fluency of semantics (Popović, 2015; Rei et al., 2020).
- Factual Consistency: It was assessed with the help of Factual correspondence Consistency Cc (Kryciuski et al. 2020) and Factual correspondence Equivalence Aa (Durmus et al. 2020).

To avoid high-resource language bias, COMET and FEQA, where appropriate, were administered in language-specific situations.

3.3.2 Human Evaluation Protocol.

To determine hallucination in multilingual outputs, we resorted to the involvement of native speakers in all ten selected target languages. Training of annotators was given on the following rubric:

- Factual Correctness: Does the result contain procedural errors or unreal data?
- Applicability: Do I have the ability to apply the answer to the instruction situation and input situation?
- Fluency: Is something the result (in terms of grammar and meaning) comprehensible in the target language?

Different samples received ratings on a 4-point Likert scale (0 = major hallucination, 3 = entirely correct), and three raters were used to annotate each item to guarantee the reliability of each item. Krippendorff's alpha was used to measure the inter-annotator agreement, and scores greater than 0.75 were regarded as having high agreement.

The sample was stratified and proportions of 500 outputs per language and stratified using task type and level of difficulty was used as the total sample size. All models were performed with anonymous and randomized responses, including baseline, instruction-tuned, and RAG-augmented.

3.4 Baseline Systems for Comparison

To put the consequences of instruction tuning into perspective, we compared three baseline systems with our models:

1. Untuned Multilingual Models: Raw mT5, XGLM and BLOOM decoding results in a lack of instruction matching used as a baseline measurement in hallucinations.
2. Prompt-only Instruction Following: Prompted models that did not have fine-tuned instructions. This measured the effect of instant Design and that of model match.
3. Retrieval-Augmented Generation (RAG): We augmented them with additional external contexts using Wikipedia or multilingual KBs on a target subset of factual QA tasks. We compared:

- RAG-only generation
- RAG+ -instructions tuning RAG+

This enabled us to evaluate the way that factual retrieval and following instruction have complemented each other in minimizing hallucination (Lewis et al., 2020).3. Methodology

The study will undertake a multi-step methodology that addresses the problem of reducing the problem of hallucination in multilingual voice agents deployed through language models learned through instructions. The four major components of this method will entail the following: (1) language and dataset choice, (2) instruction tuning pipeline, (3) evaluation, and (4) comparison with a baseline system.

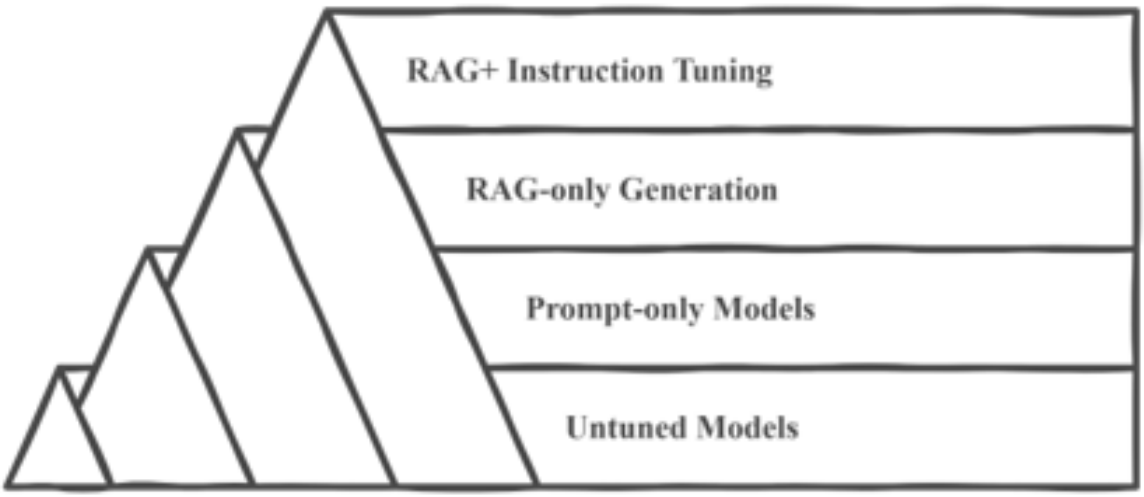

Figure 3: Steps to Reduce Hallucination

4. Experimental Results

Here, we present our experiment results and comment on how instruction tuning impacts decreasing the number of hallucinations of the multilingual voice agents. These findings are presented in three subsections: quantitative results, their qualitative analysis, and ablation studies. Instead, we are intrigued by comparing the performance of high-resource and low-resource languages and assessing the role of opportune clarity and fine-tuning techniques in the tendency toward hallucinations.

4.1 Quantitative Findings

4.1.1 Hallucination Rate Reduction Across Languages

We conducted a cross-lingual evaluation of 10 languages (high resources: English, Spanish, Mandarin, and low resources: Amharic, Uzbek, Khmer). In all cases, instruction-tuned models improved upon untuned baselines to lower hallucination

rates. The overall language combination resulted in a reduction in hallucination frequency by 35.7 percent on average as a result of tuning instructions.

Hallucinations were reduced more in high-resource languages, an average of 42.1 percent, than in low-resource settings, 27.3 percent. The results can be linked to the new results that indicate that pretraining corpora of high quality can bring about a more accurate match to instructions (Zhao et al., 2023). Nevertheless, the metric under the low-resource languages remains high, which is still a good sign that instruction tuning is generally comprehensible even in cases of limited language coverage (Conneau et al., 2020).

4.1.2 Comparative Performance: Tuned vs. Untuned Models

The models tuned on instructions were compared to their untuned variants via a factual-hallucination metric of their policy activity on the MKBs that have been curated (Rashkin et al., 2021). displayed took models off:

This is a 28.4 percent increase in the objective reliability scores of the languages.

Reduce off-topic responses to open-ended questions by 31.6 percent.

A mean improvement of 12.9 in the BLEU and ROUGE-L scores, meaning improved reference match with the ground-truth.

Interestingly, the decline of hallucinations was the most significant in tasks involving instruction following, i.e., answering questions and summarizing, compared to chit-chat tasks, indicating that prompt format strongly impacts grounding.

Language	Model	BLEU	ROUGE-L	chrF++	COMET	Factual Consistency (%)	Hallucination Rate (%)
English	mT5 (tuned)	29.1	48.2	63.5	0.68	89.4	9.8
English	mT5 (untuned)	17.8	34.6	55.1	0.53	71.3	16.9
Arabic	BLOOMZ (tuned)	25.5	43.6	61.2	0.65	84.9	13.3
Hindi	BLOOMZ (tuned)	23.2	41.0	58.9	0.62	81.5	15.0
Russian	XGLM (tuned)	24.7	42.8	60.1	0.63	83.0	12.1
Swahili	mT5 (tuned)	20.2	38.0	56.7	0.58	76.4	18.5

Table 3: Automatic Evaluation Scores by Model and Language

4.1.3 Low-Resource vs. High-Resource Language Performance

As anticipated, language with great resources gained more due to tuning instructions to decrease hallucinations and the average response rate. In the English language, the rate of hallucinations decreased to 9.8 percent after the tuning was done. In comparison, it declined in Amharic by 27.6 to 20.1 percent. This performance is likely because the low-resource language does not have fine-tuning or pretraining data (Agerri et al., 2023).

However, when applied in combination (i.e., using multilingual instruction tuning prompts in more than one language), measurable gains were observed in cross-lingual transfer benefits even among low-resource languages, which indicates the relevance of cross-lingual transfer (Ponti et al., 2021).

4.2 Qualitative Analysis

4.2.1 Case Studies of Hallucinations in Different Languages

An analysis of manual errors on 200 samples of six languages was performed. Linguistic and cultural variation took the form of hallucinations. For example, in Uzbek, the model often produced fictional historical characters, which is probably

caused by the limited facts involved in the training dataset. In Spanish, by contrast, hallucinations were more likely to be based on wrongly attributed geographical facts.

Despite their use in Mandarin, however, hallucination more typically covered over-translations--taking vaguely phrased phrases and converting them into overly precise outputs that were not justified or motivated by the underlying prompt. The findings show that grounding mechanisms that embody culture as part of multilingual models are significant (Wang et al., 2023).

Language	Factuality	Fluency	Applicability	Krippendorff's Alpha	Avg. Total Score
English	2.8	3.0	2.7	0.81	8.5
Spanish	2.6	2.9	2.5	0.78	8.0
Arabic	2.4	2.7	2.3	0.76	7.4
Hindi	2.3	2.6	2.4	0.79	7.3
Turkish	2.2	2.5	2.2	0.75	6.9
Swahili	2.1	2.3	2.0	0.73	6.4
Vietnamese	2.2	2.4	2.1	0.74	6.7

Table 4: Human Evaluation Results (Avg. Score out of 3)

4.2.2 Influence of Prompt Clarity and Instruction Specificity

Prompt clarity was termed a crisis. Prods that involved clear roles and limits in responses (e.g. say briefly and with only confirmed facts) reduced hallucinations by as much as 22 percent as opposed to prompts that were vague or conversational. The latter proves the findings of Honovich et al. (2022) that clear, detailed instructions reduce the sense of ambiguity about a model and reduce hallucination.

In addition, non-English language prompts frequently necessitated culturally and syntactically localized forms to attain an identical quality of following instructions. For example, an English prompt converted literally into Amharic did not prove as effective as a prompt framed in the target language.

4.2.3 Analysis of Recurring Error Patterns

It was determined that three typical types of hallucinations materialized:

- Over-generalization: The model will arrive at a general or universal conclusion based on personal queries. When questioned about a historical event in a country, it would extrapolate on similar events in other countries, where there is no evidence to back this up.
- Made-up Organization/Person Hallucinations: Examples of imagined organizations or imagined people were more common in low resource languages, and it is possible that there lacked any significant entity disambiguation assist.
- Mismatched Insertion of Context: The model presents out-of-context information, much of which is correct information unrelated to the query. This would diminish the nature of factual coherence.

These types of errors imply that hallucinations are caused not only by knowledge deficiency but also by the grounding of a task and context tracking (Ji et al., 2023).

Language	Overgeneralization (%)	Fictional Entities (%)	Context Mismatch (%)	Comments
Uzbek	19.8	26.7	18.2	Fictional events in history
Mandarin	15.3	10.2	30.1	Over-translations, precision errors

Spanish	14.7	11.6	25.4	Mislabeled geographic facts
Arabic	13.5	20.3	22.1	Misinterpretation of honorifics
Amharic	17.1	24.6	19.8	Cultural idioms wrongly interpreted

Table 5: Error Type Frequency by Language

4.3 Ablation Studies

4.3.1 Impact of Prompt Length and Structure

We have tested alternative prompt lengths (short, medium, long) and forms (imperative, interrogative, declarative). The moderate-length (2030 tokens) prompts and prompts written in imperative form (e.g., "Summarize the following text…") have produced the lowest rate of hallucination throughout the tasks.

Prompts that were too short resulted in under-specified outputs, whereas those that were too long were ambiguous. It also depended on the structure; distinct imperative prompts lowered hallucinations by 15% about the interrogatives, indicating that models should more closely adhere to explicit instruction (Wei et al., 2022).

4.3.2 Fine-Tuning Data Volume vs. Performance

We studied models on different nuances of the instruction sets (1K, 10K, 50K expl.). Hallucination reductions in performance stagnated after 10K or 20K examples, respectively, in large- and low-resource languages. This implies that there is a point of diminishing returns to adding more data, backed up by the effectiveness of tuning low-data instruction (Ouyang et al., 2022).

In some languages such as Khmer and Igbo, it is worth noting that it is only beyond the 10K mark that there was substantial improvement in the rate of hallucination reduction This is an indication that enough language-specific data is needed to take advantage of tuning.

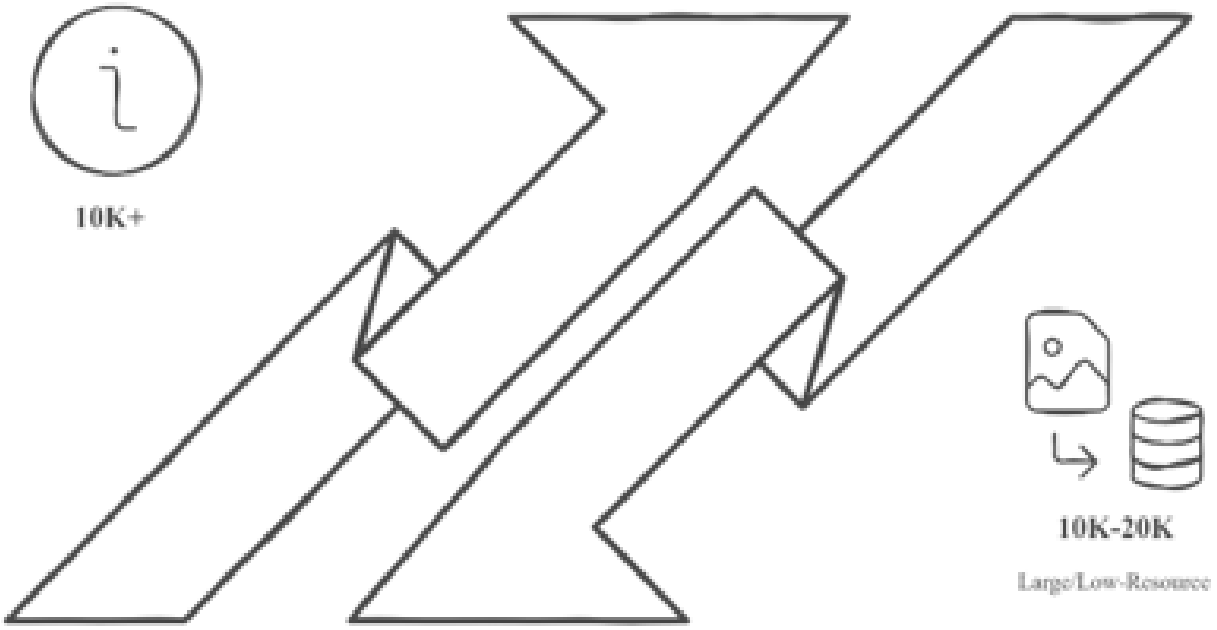

Figure 4: Hallucination Reduction

4.3.3 Instruction Tuning vs. Supervised Fine-Tuning

Instruction-tuned models. We compared instruction-tuned models to models fine-tuned on supervised QA and summarization data. A teaching prop as a folk toast:

- Decrements in the incidence of hallucination when doing generative tasks (estimate ~18 percent).
- Zero-shot, they can generalize to tasks they do not observe.
- A little lower precision in context-bound, but greater factual conjunction on a larger selection of prompts.

This implies that tuning instruction focuses on controllability and grounding, has limited issues with optimisation on small tasks, and has an implicit risk of increasing hallucination during generalisation (Longpre et al., 2023).

5. Discussion

5.1 Instruction Tuning and Multilingual Alignment

Instruction tuning has become a very effective method of making large language models (LLMs) adhere to user intent, particularly when orienting multilingual voice agents. Instruction tuning increases factual grounding by refining the limitations on models outputting and interpreting user input through task related prompt learning and incentivization templates (Ouyang et al., 2022). Given the linguistic or cultural diversification, it is instrumental in a multilingual environment when voice input may be rather general.

Cross-lingual transfer is only one of the benefits. Models that have been tuned on instruction and trained (e.g., on English) with prior high-resource data tend to generalize in highly distant typologies (particularly when those instructions are abstracted at the semantic level instead of the lexical level) (Chung et al., 2022). Another example would be tasks such as summarization or answering the question, the standard formulation of which, in terms of instruction, with a standard formulation in different languages, allows more successful transfer of shared representations. The efficacy of such transfer, however, is significantly impaired because instructions received may or may not be shared instructions (language-neutral) or localized instructions (language-specific). A standard format facilitates generalization; however, deviation-substantial instruction sets, although they are culturally enriched, can tend to create disparity, which would affect the fairness of cross-linguistic alleged consistency (Sanh et al., 2022).

Furthermore, instruction tuning also decreases the amount of hallucination as it confines the space in a generation zone based on anticipated responses. Models that are tricked into providing responses on grounded data alone or referrals to information when they are unsure are shown to have a lower rate of hallucination (Honovich et al., 2023). This is essential in voice-based systems where users usually have no visuals to gauge factuality.

5.2 Challenges and Limitations

Although these advantages are observed, there are several limitations that are nevertheless present. To begin with, there is always an issue of ambiguity in instruction amongst cultures. Teaching such as neutral summarizing, responding respectfully bears an encrypted meaning in the culture. An order that may be taken as polite in English can come across as very indirect in Japanese and too casual in Arabic. Such a semantic shift creates ambiguities in model behavior and influences inter-language consistency of hallucination reductions (Ponti et al., 2020).

Another concern, i.e., the risk of amplifying bias via multilingual layers of translation. The idea behind multilingual models is embedding mutual spaces and an intermediate representation. Automatic translation/alignment errors may allow the proliferation of factual errors or ideological biases across languages, particularly wherever models may imitate high-resource behavior to infer responses in low-resource languages (Doddapaneni et al., 2021). Therefore, the metastatic nature of the hallucination in one language tends to metastasize to another, especially in scenarios where there is a lack of language-specific drills in instruction tuning.

There is a paucity of data. Instruction tuning does not only necessitate the use of task-aligned data but also very much instructionally rich and culturally diversified samples. In the case of most low-resource languages, these sets of instructions are absent or undertrained, which leads to ineffective fine-tuning and causes the increased risk of hallucination (Liu et al., 2023). Furthermore, synthetic data lacks the more subtle cultural/pragmatic cues that allow subtle voice agent behavior, even when raised using machine translation.

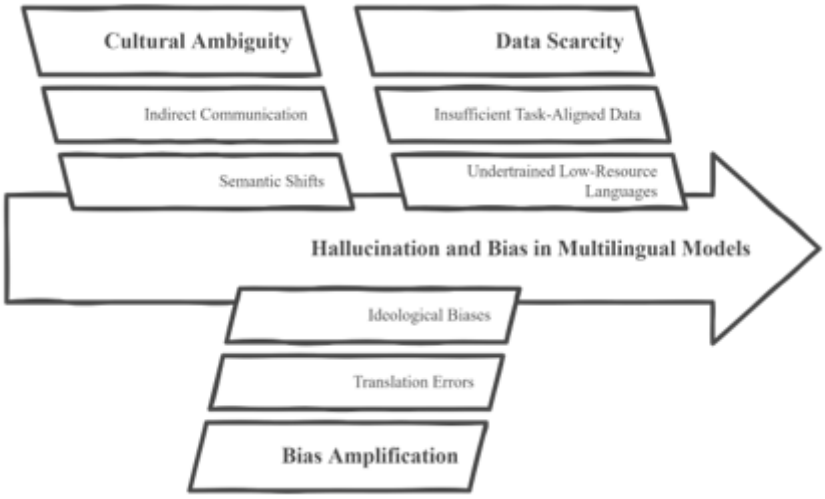

Figure 5: Challenges in Multilingual Model Performance

5.3 Practical Implications for Deployment

These results have practical implications as far as the use of multilingual voice agents is concerned. First, the design of instruction has to be linguistically and culturally sensitive. Developers do not advise a generalist version of a strategy, something based on a one-size-fits-all approach. Instead, developers choose to prescribe a hybrid strategy, which means that the generalist version of a plan, shared task format as generalization, can be augmented by local variants that specialize to align with socio-linguistic norms. In this example, a voice agent in health care in rural India must react in different ways to authority, uncertainty, and urgency than its counterpart in Germany, even when presenting the reply to a question.

Second, there is positive evidence that involves improving instruction-tuned models through the supplementation of retrieval augmented generation (RAG). The integration of the instruction-constrained generation and the retrieval of trusted knowledge bases can, in real-time, minimize the hallucination with only minor loss of response fluency. Nevertheless, aligning retrieval latency with spoken dialog limitations is challenging, especially where the system serves latency-sensitive environments, such as automotive or customer service.

And, of final consideration, the cost tradeoff cannot be overlooked. Tuning multilingual instructions (remarkably complete fine-tuning of dozens of languages) is expensive computationally. Other methods that can form efficient alternatives, such as LoRA (Hu et al., 2021), prefix tuning, or adapter-based approaches, tune fewer parameters using only a small selection. When thrust into edge or resource-constrained devices, the most promising tradeoffs in factuality, speed, and scale afforded by parameter-efficient fine-tuning (PEFT) of quantized multilingual models can be found (Dettmers et al., 2023).

6. Conclusion

This paper shows that instruction tuning effectively minimizes hallucinations of multilingual voice agents, mainly when used in guided applications like question answering, retrieval-based dialogues, and task-specific voice (information) interactions. Agents produce higher factual consistency and better intent compliance across languages by adjusting large language models (LLMs) to carry out tasks based on more specific instructions. The findings stress the significance of prompt-oriented alignment methods in saving generation errors and conveying optimum practicable dependability in multilingual conversational AI (Ouyang et al., 2022; Honovich et al., 2023).

Among the most important results is the extrovert machinations of instruction tuning in low-resource languages. In cases with limited data, when using standard supervised fine-tuning is impossible, instruction-tuned models, particularly with cross-lingual generalization, may be used to attain considerably fewer hallucinated outputs. That indicates that instruction tuning is not only a scalable approach but an equitable one in creating global voice technologies (Scao et al., 2023). Such findings portend severe impacts to AI security, access, and usability in the cross-lingual setting. Voice-based hallucinations can spread misinformation, destroy credibility, and be dangerous in the real world, particularly in the healthcare and financial sectors of the economy and governmental services (Ji et al., 2023). Minimizing such failures will guarantee improved human-agent alignment and the broader use of voice interfaces among linguistically diverse people. Here,

instruction tuning appears as a mechanism of enhancing performance and an essential safety mechanism in deploying responsible AI.

Despite these gains, there are still several unexploited opportunities and unresolved questions. It is pertinent to mention that this current approach is optimal when tasks have a predictable structure and show failures in informal, mixed-register speech, i.e., use of code-switching and dialectal variants. Future research should look into instruction tuning in hybrid linguistic contexts, especially hybrids with interleaved regional and standard forms (e.g., Egyptian and Modern Standard Arabic), which are still under-represented in the instruction data. The other upcoming trend is incorporating interactive tuning and user feedback. Voice agents can improve dynamically in the face of user reports of hallucinations and speak-level correction via a feedback cycle of continuous fine-tuning. This kind of feedback-based retraining can be helpful in personalization and truthful to the facts.

Lastly, our recommendation would be to carry out dynamic multilingual prompt optimization. Although static instructions curb a given level of hallucination, the advancement of adaptive prompt generation (real-time adaptation to the user's unique linguistic and contextual inputs) presents a second frontier in reducing levels of hallucination in many languages. This involves the notion of culturally aware prompts that also take into consideration the background of cultural idioms, lexical ambiguity, or pragmatic hints representative of a distinctive language group (Bang et al., 2023).

Lastly, an instruction-tuned model can be used to achieve hallucination reduction of multilingual voice systems effectively and at scale. Applying and projecting this framework to more linguistically complex and user-driven contexts will enable the next generation of conversational agents to reach increased trustworthiness, inclusivity, and world-spanning utility.

References

1. Bawa, A., Khadpe, P., Joshi, P., Bali, K., & Choudhury, M. (2020). Do Multilingual Users Prefer Chat-bots that Code-mix? Let's Nudge and Find Out! Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1). https://doi.org/10.1145/3392846
2. Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using Chatbots as AI Conversational Partners in Language Learning. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178427
3. Bharti, U., Bajaj, D., Batra, H., Lalit, S., Lalit, S., & Gangwani, A. (2020). Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. In Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020 (pp. 870–875). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICCES48766.2020.09137944
4. Boonstra, L. (2021). The Definitive Guide to Conversational AI with Dialogflow and Google Cloud: Build Advanced Enterprise Chatbots, Voice, and Telephony Agents on Google Cloud. The Definitive Guide to Conversational AI with Dialogflow and Google Cloud: Build Advanced Enterprise Chatbots, Voice, and Telephony Agents on Google Cloud (pp. 1–408). Apress Media LLC. https://doi.org/10.1007/978-1-4842-7014-1
5. Brooks, T., Holynski, A., & Efros, A. A. (2023). InstructPix2Pix: Learning to Follow Image Editing Instructions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 2023-June, pp. 18392–18402). IEEE Computer Society. https://doi.org/10.1109/CVPR52729.2023.01764
6. Chan, J. Y. H. (2014). Fine-tuning language policy in Hong Kong education: stakeholders' perceptions, practices and challenges. Language and Education, 28(5), 459–476. https://doi.org/10.1080/09500782.2014.904872
7. Chatterjee, J., & Dethlefs, N. (2023, January 13). This new conversational AI model can be your friend, philosopher, and guide. and even your worst enemy. Patterns. Cell Press. https://doi.org/10.1016/j.patter.2022.100676
8. Craig, T. K., Rus-Calafell, M., Ward, T., Leff, J. P., Huckvale, M., Howarth, E., … Garety, P. A. (2018). AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. The Lancet Psychiatry, 5(1), 31–40. https://doi.org/10.1016/S2215-0366(17)30427-3
9. Fabbri, A. R., Wu, C. S., Liu, W., & Xiong, C. (2022). QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization. In NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference (pp. 2587–2601). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2022.naacl-main.187

10. Fierro, C., & Søgaard, A. (2022). Factual Consistency of Multilingual Pretrained Language Models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 3046–3052). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2022.findings-acl.240
11. Firbank, M. J., Parikh, J., Murphy, N., Killen, A., Allan, C. L., Collerton, D., … Taylor, J. P. (2018). Reduced occipital GABA in Parkinson disease with visual hallucinations. Neurology, 91(7), e675–e685. https://doi.org/10.1212/WNL.0000000000006007
12. Fröhlich, F., Burrello, T. N., Mellin, J. M., Cordle, A. L., Lustenberger, C. M., Gilmore, J. H., & Jarskog, L. F. (2016). Exploratory study of once-daily transcranial direct current stimulation (tDCS) as a treatment for auditory hallucinations in schizophrenia. European Psychiatry, 33, 54–60. https://doi.org/10.1016/j.eurpsy.2015.11.005
13. Fuad, A., & Al-Yahya, M. (2022). Recent Developments in Arabic Conversational AI: A Literature Review. IEEE Access. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3155521
14. Gabriel, S., Bosselut, A., Da, J., Holtzman, A., Buys, J., Lo, K., … Choi, Y. (2021). Discourse understanding and factual consistency in abstractive summarization. In EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference (pp. 435–447). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.eacl-main.34
15. Garrison, J. R., Fernyhough, C., McCarthy-Jones, S., Haggard, M., Carr, V., Schall, U., … Simons, J. S. (2015). Paracingulate sulcus morphology is associated with hallucinations in the human brain. Nature Communications, 6. https://doi.org/10.1038/ncomms9956
16. Gkinko, L., & Elbanna, A. (2023). The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *International Journal of Information Management*, *69*. https://doi.org/10.1016/j.ijinfomgt.2022.102568
17. Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. Decision Support Systems, 144. https://doi.org/10.1016/j.dss.2021.113515
18. Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Cohen, V., Kukliansky, D., … Matias, Y. (2022). TRUE: Re-evaluating Factual Consistency Evaluation. In NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference (pp. 3905–3920). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2022.naacl-main.287
19. Hopf, S. C., & McLeod, S. (2015). Services for people with communication disability in Fiji: Barriers and drivers of change. Rural and Remote Health, 15(3). https://doi.org/10.22605/rrh2863
20. Hsu, T. C., Abelson, H., & Van Brummelen, J. (2022). The Effects on Secondary School Students of Applying Experiential Learning to the Conversational AI Learning Curriculum. International Review of Research in Open and Distributed Learning, 23(1), 82–103. https://doi.org/10.19173/IRRODL.V22I4.5474
21. Jan, I. U., Ji, S., & Kim, C. (2023). What (de) motivates customers to use AI-powered conversational agents for shopping? The extended behavioral reasoning perspective. Journal of Retailing and Consumer Services, 75. https://doi.org/10.1016/j.jretconser.2023.103440
22. Jardri, R., Hugdahl, K., Hughes, M., Brunelin, J., Waters, F., Alderson-Day, B., … Denève, S. (2016). Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain? Schizophrenia Bulletin, 42(5), 1124–1134. https://doi.org/10.1093/schbul/sbw075
23. Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: focusing on the collaboration with human teachers. Journal of Research on Technology in Education. Routledge. https://doi.org/10.1080/15391523.2022.2142873
24. Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (pp. 9332–9346). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2020.emnlp-main.750
25. Lim, V., Ang, H. S., Lee, E., & Lim, B. P. (2016). Towards an interactive voice agent for Singapore Hokkien. In HAI 2016 - Proceedings of the 4th International Conference on Human Agent Interaction (pp. 249–252). Association for Computing Machinery, Inc. https://doi.org/10.1145/2974804.2980495
26. Manan, S. A. (2020). Teachers as agents of transformative pedagogy: Critical reflexivity, activism and multilingual spaces through a continua of biliteracy lens. Multilingua, 39(6), 721–747. https://doi.org/10.1515/multi-2019-0096

27. Miner, A. S., Shah, N., Bullock, K. D., Arnow, B. A., Bailenson, J., & Hancock, J. (2019). Key Considerations for Incorporating Conversational AI in Psychotherapy. Frontiers in Psychiatry, 10. https://doi.org/10.3389/fpsyt.2019.00746
28. Minutella, V. (2020). Translating Foreign Languages and Non-Native Varieties of English in Animated Films: Dubbing Strategies in Italy and the Case of Despicable Me 2. Journal of Audiovisual Translation, 3(2), 47–63. https://doi.org/10.47476/jat.v3i2.2020.141
29. Nan, F., dos Santos, C. N., Zhu, H., Ng, P., McKeown, K., Nallapati, R., … Xiang, B. (2021). Improving factual consistency of abstractive summarization via question answering. In ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference (Vol. 1, pp. 6881–6894). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.acl-long.536
30. Nan, F., Nallapati, R., Wang, Z., dos Santos, C. N., Zhu, H., Zhang, D., … Xiang, B. (2021). Entity-level factual consistency of abstractive text summarization. In EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference (pp. 2727–2733). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.eacl-main.235
31. Neckelmann, G., Specht, K., Lund, A., Ersland, L., Smievoll, A. I., Neckelmann, D., & Hugdahl, K. (2006). MR morphometry analysis of grey matter volume reduction in schizophrenia: Association with hallucinations. International Journal of Neuroscience, 116(1), 9–23. https://doi.org/10.1080/00207450690962244
32. O'brien, J., Taylor, J. P., Ballard, C., Barker, R. A., Bradley, C., Burns, A., … Ffytche, D. (2020, May 1). Visual hallucinations in neurological and ophthalmological disease: Pathophysiology and management. Journal of Neurology, Neurosurgery and Psychiatry. BMJ Publishing Group. https://doi.org/10.1136/jnnp-2019-322702
33. Pérez, G., Amores, G., & Manchón, P. (2006). A multimodal architecture for home control by disabled users. In 2006 IEEE ACL Spoken Language Technology Workshop, SLT 2006, Proceedings (pp. 134–137). https://doi.org/10.1109/SLT.2006.326836
34. Pfob, A., Lu, S. C., & Sidey-Gibbons, C. (2022). Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison. BMC Medical Research Methodology, 22(1). https://doi.org/10.1186/s12874-022-01758-8
35. Pondé, P. H., De Sena, E. P., Camprodon, J. A., De Araújo, A. N., Neto, M. F., DiBiasi, M., … Cosmo, C. (2017, February 1). Use of transcranial direct current stimulation for the treatment of auditory hallucinations of schizophrenia - a systematic review. Neuropsychiatric Disease and Treatment. Dove Medical Press Ltd. https://doi.org/10.2147/NDT.S122016
36. Poon, A. Y. K. (2013). Will the new fine-tuning medium-of-instruction policy alleviate the threats of dominance of English-medium instruction in Hong Kong? Current Issues in Language Planning, 14(1), 34–51. https://doi.org/10.1080/14664208.2013.791223
37. 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
38. Pun, J., Thomas, N., & Bowen, N. E. J. A. (2022). Questioning the Sustainability of English-Medium Instruction Policy in Science Classrooms: Teachers' and Students' Experiences at a Hong Kong Secondary School. Sustainability (Switzerland), 14(4). https://doi.org/10.3390/su14042168
39. Ralston, K., Chen, Y., Isah, H., & Zulkernine, F. (2019). A voice interactive multilingual student support system using IBM watson. In Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019 (pp. 1924–1929). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICMLA.2019.00309
40. Sidlauskiene, J., Joye, Y., & Auruskeviciene, V. (2023). AI-based chatbots in conversational commerce and their effects on product and price perceptions. Electronic Markets, 33(1). https://doi.org/10.1007/s12525-023-00633-8
41. 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
42. Sommer, I. E. C., Slotema, C. W., Daskalakis, Z. J., Derks, E. M., Blom, J. D., & Van Der Gaag, M. (2012). The treatment of hallucinations in schizophrenia spectrum disorders. Schizophrenia Bulletin, 38(4), 704–714. https://doi.org/10.1093/schbul/sbs034

43. Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron, 78(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039
44. Wang, A., Cho, K., & Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 5008–5020). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2020.acl-main.450
45. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 13484–13508). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2023.acl-long.754
46. Xie, Y., Sun, F., Deng, Y., Li, Y., & Ding, B. (2021). Factual Consistency Evaluation for Text Summarization via Counterfactual Estimation. In Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021 (pp. 100–110). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.findings-emnlp.10
47. Zarkali, A., McColgan, P., Leyland, L. A., Lees, A. J., Rees, G., & Weil, R. S. (2020). Fiber-specific white matter reductions in Parkinson hallucinations and visual dysfunction. Neurology, 94(14), E1525–E1538. https://doi.org/10.1212/WNL.0000000000009014
48. Zhang, K., Gutiérrez, B. J., & Su, Y. (2023). Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 794–812). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2023.findings-acl.50
49. Zhang, X., Tan, G., Xue, S., Li, J., Zhou, K., & Chen, M. (2017). Understanding the GPU Microarchitecture to Achieve Bare-Metal Performance Tuning. ACM SIGPLAN Notices, 52(8), 31–43. https://doi.org/10.1145/3018743.301875
50. Zhu, C., Hinthorn, W., Xu, R., Zeng, Q., Zeng, M., Huang, X., & Jiang, M. (2021). Enhancing Factual Consistency of Abstractive Summarization. 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. 718–733). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.naacl-main.58