

Putting Audits on the Bench: An Algorithm Audit Informed Philosophy to Construct Benchmarks for LLMs using Low Resource Language Dialects for Direct Comparison

William Coggins
william.coggins@ufl.edu

September 20, 2026

Contents

ACKNOWLEDGMENTS	4
LIST OF TABLES	5
LIST OF FIGURES	6
LIST OF ABBREVIATIONS	8
LIST OF TERMS	9
1 ABSTRACT	10
2 INTRODUCTION	10
2.1 Research Objectives	11
2.1.1 Evaluate Algorithm Auditor Methodologies	11
2.1.2 Evaluate LLMs on the basis of traits that leak information	11
2.1.3 Create Tools to evaluate LLMs on relative performance	12
2.2 Expected Contributions	12
3 LITERATURE REVIEW	13
3.1 Algorithm Auditing	13
3.2 NLP & Low Resource Languages	14
3.3 LLM Benchmarks	15
3.4 User Populations for Models versus Training Data and Auditing for Generative Models	15
4 Proposed Work: Interview Study on Algorithm Auditors	16
4.1 Background	16
4.2 Methodology	17
4.2.1 Participants	17
4.2.2 Interview Methodology	17
4.2.3 Question Design and Goals	18
4.2.4 Analysis	18
4.3 Results	18
4.3.1 Results: Interview Profiles	18
4.3.2 Results: Interview Codes	18
4.4 Discussion	19
4.4.1 No Set Standards	19
4.4.2 Stakeholder Involvement	19
4.4.3 Subject Matter Experts	20
4.4.4 Context Sensitivity	20
4.4.5 Security	20
4.4.6 Top-down vs Bottom-up: A Philosophical Approach	20
4.4.7 Policy-Motivated Audit vs. Voluntary Audit	21
4.5 Conclusion	21
5 Proposed Work: Testing LLMs on QA in English Dialects	22
5.1 Background & Related Work	22
5.1.1 Linguistic Background	22
5.1.2 Related Work	23
5.2 Methodology	23
5.2.1 Study Design	23
5.2.2 LLM Selection	23
5.2.3 Data Collection	24
5.2.4 Question Dialect Translation	24
5.2.5 Evaluation	25

5.2.6	Analysis Design	25
5.3	Results	25
5.3.1	LLM Performance	25
5.3.2	Dialect Results	25
5.4	Discussion	27
5.4.1	Same Question: Different Answer	27
5.4.2	Responses Mirroring Dialects	28
5.5	Proposed Extended Work	28
6	Proposed Work: A Benchmark for Dialect Comparison (DiaBench)	29
6.1	Related Work	29
6.2	DiaBench Description	29
6.2.1	Dialect Categories	30
6.2.2	Language Comprehension Category	30
6.2.3	Reasoning Category	30
6.2.4	QA Category	30
6.2.5	Safety Category	30
6.2.6	Contamination Efforts	31
6.3	Experiments	31
6.3.1	Models	31
6.3.2	Training and Evaluation	31
6.4	Meaningful Contribution	32
7	Tentative Timeline	33
	LIST OF REFERENCES	34
	APPENDIX	40

ACKNOWLEDGMENTS

LIST OF TABLES

1	Interviewee profiles and notable audits conducted	19
2	Examples of questions that were asked to participants concerning conducted audits	19
3	Example SAE question with WAPE and AAVE equivalents	24
4	SAE Questions and LLM responses with correctness evaluation: Correct (1) / Incorrect (0)	25
5	Binomial Tests comparing dialect errors in LLMs to corresponding SAE performance for same LLM. $N = 100$, $\alpha = .05$	27
6	Dialect questions with different response types	28
7	Benchmark Categories and Score Percentage	29
8	Benchmark Categories and Example Questions	30
9	List of LLMs that will be experimented on for DiaBench validation.	31

LIST OF FIGURES

1	Dialect Translation and LLM prompting	23
2	Methodology Flowchart	24
3	Errors produced across dialects grouped by LLM	26
4	Errors produced across LLMs	26
5	Errors produced across dialects	26
6	Interview Questions and Procedure	42

LIST OF OBJECTS

LIST OF ABBREVIATIONS

AAVE	AAVE refers to African American Vernacular English, the dialect originating out of the combination of Southern English and slang from enslaved Africans in the American South.
LLM	LLM refers to Large Language Model, which are a category of deep learning models capable of understanding and generating natural language.
QA	QA refers to Question and Answer. This is a category used in benchmarks where a model is tasked with generating a response to a question, specifically with regards to information retrieval as commonly represented by questions such as trivia.
SAE	SAE refers to Standard American English, the standardized variety of English used in the United States for formal writing, and communication. It is characterized by certain rules of spelling, grammar, punctuation, and vocabulary that are promoted as the norm for discourse.
WAPE	The proposal makes frequent reference to WAPE. WAPE refers to West African Pidgin English, the English dialect spoken across West Africa and areas of its modern diaspora including Western Europe and the Americas.

LIST OF TERMS

Bias

While there are many definitions of bias, and it is a pertinent term in discussions of statistics and machine learning specifically, we use bias in this document to refer to differences in treatment and subsequent outcome largely attributed to the basis of traits such as race, gender, and socioeconomic status. These factors discuss characteristics of people that ultimately should not significantly factor into a decision.

Fairness

While there are many definitions of fairness, we use the term to generally mean the ability for models to generate impartial outcomes regardless of the protected characteristics like race or age.

1 ABSTRACT

Widely-used models such as LLMs have a large population of users. I study the effect of dialects on the accuracy of these LLMs since a large population of users often do not have the same exact textual input usually mirroring their particular dialect as represented in text. Early results indicate that some dialects can markedly reduce the efficacy of these models, and as such, they should be evaluated closely. Since these models can lead to harmful outputs for users and for implementers, they are included among the many models that are audited. Algorithm auditors are key individuals who work with developers in a holistic way in order to identify concerns from different parties, validate those concerns through testing, and attempt to provide recommendations and remediation for implementers. I interview these auditors to discover trends in their practices as well as commonalities as the profession becomes a growing necessity among businesses and governments implementing AI across their organizations. In order to democratize testing these models with regards to dialects in a post-hoc fashion, I plan to develop DiaBench. DiaBench is an LLM benchmark tool that, for each english prompt, a functionally equivalent english dialect prompt will be generated and processed by the LLM for analysis in discrepancies in interpretation. This tool aids auditors in verifying the integrity of large language models and their ability to accurately comprehend prompts from various user groups. Each of these efforts extend toward making these large language models more accessible by many user groups worldwide, both by making aware these discrepancies exist, as well as a solution to help those who audit these models with identifying disparities of usability within models easily.

2 INTRODUCTION

As machine learning has established itself as the dominant method to create artificial intelligence systems, it bears both unique benefits and harms to its application. As it acquires knowledge through the set of training data and attempts to generalize and as a result, its functionality is in part limited to the range of training data available. As apparent in virtually every machine learning system, there will be gaps or discrepancies between the data provided to train the system and the data present in real life. A common reason for this gap can be that there is insufficient data for these systems to readily train on. A lack of data can be an issue of capture, where it is hard to get this information, be it structural where systems are not set up to capture this information, or it could be based on population, where there is too sparse an existence of data of a particular group that makes it hard to capture in a natural environment. While methods like data imputation and synthetic data generation help assuage these issues, these gaps have to be first identified in order for imputation to occur. These discrepancies can produce inaccuracies in the outputs of these models. While these inaccurate outputs can be innocuous, they can be substantially harmful depending on the application of the system. When applied specifically to Large Language Models (LLMs), the discrepancies can be two-fold: in the area where the query posits information, and the actual input the user prompts the system with. In the former case, the model does not have the information such that retrieval or attempting to answer the query will likely result in hallucinations to occur. In the case of the latter, the input the user prompts the system with can be formatted in such a way that increases the ambiguity of the prompt, thus inhibiting the model’s ability to successfully answer it in part due to how much training data the model has on the input. If the input features typos or errors, then models have a higher propensity to incorrectly interpret the input prompt [86]. While developers attempt to fix these issues, the process of a model’s development and deployment often involve more than just the developers themselves, and have to administer some governance structure especially if these models may be subject to regulation or policies both internal and external. In order to assess both these models and governance structures, algorithm auditors are often brought in to provide a thorough analysis of these systems.

Algorithm auditors are a burgeoning profession of individuals who inspect algorithms for harmful behavior. These audits can include analysis of both systems end to end as well as the personnel and governance structures surrounding the implemented system, making it a wholistic approach towards improving the model [38]. Trends indicate that usage of AI systems by businesses extended toward integration with LLMs for employees to use as well as for customers [71]. This then follows that with regards to LLMs specifically, auditing efforts must be created to match the demand. As LLM usage is utilized both internally by employees and contractors, and externally by general audiences, there is reasonable concern for harmful or inaccurate output. Harmful output has the capacity to change the interaction for external customers of the business such that information provided does not match articulated policy [68]. Inaccurate output internally can reduce adoption of said models since they degrade user trust. Through a study of LLMs using dialect-based approaches, I and several fellow researchers found that functionally equivalent prompts in different dialects had a higher propensity to generate false or harmful output compared to prompts in Standard American English [16]. In order for businesses to widen the scope of their customers as well as their internal developers, auditors need tools that are able to address this specific facet of LLM development.

Algorithm auditing as a practice has several frameworks created by a multitude of institutions. These institutions can sometimes be governments as demonstrated by NIST’s AI risk management framework [72], or the numerous frameworks spawned in response to EU’s AI Act [25] [28] [53]. Other institutions can include private companies such as those that

specialize in building models such as Microsoft and their Responsible AI Tool Suite¹, Google's Secure AI Framework², or auditing firms which specialize in the practice such as Babl AI's Assurance audit framework [44]. These frameworks, when they are low-level enough to include steps towards practical assessments, often include information related to benchmarks that models can aim to target for what can be considered a functioning model. These benchmarks as tools are extremely helpful in companies' ability to assess their own or other capabilities with regards to interested criteria. Current benchmarks with regards to LLMs look at the LLMs' ability to reason, produce correct code generation, play games, and any number of tasks related to what users call upon LLMs to perform accurately [76] [80]. These benchmarks however evaluate the LLMs' ability to complete these tasks assuming the input is likewise similar, that being prompts that are formatted with standard english primarily. Attempts to look at language largely inspect a LLMs' ability to comprehend language with regards to fixing issues, as opposed to having the robustness to comprehend a larger swathe of words and grammar rules that dialects often bear. This creates a gap of evaluation where users do not have a sufficient benchmark to evaluate if their input provides a similar output to other users even when those inputs differ not by task, but by the dialect the task is being prompted in. There then becomes a significant need for this type of evaluation as user groups expand past those who primarily utilize standard english as their mode of interaction. This idea of the benchmark becomes relevant as a critical piece of my dissertation. Three pieces comprise the composition of my proposal:

1. An interview study conducted with practicing algorithm auditors.
2. An experiment concerning dialects being used as inputs for large language models.
3. A software benchmark for dialect-based experimentation for prompts purposed for LLMs.

These three pieces are undergirded by a few ideas. Firstly, that computing systems should aim to be accessible and ideally equitable. Equitable in this instance refers to having the propensity to lack bias common in systems. In an applied context, equity amounts to improving algorithms. Those who may be best equipped to perform such a feat are auditors, as they help improve algorithms. Improving tools and benchmarks for auditors gives auditors greater ability to improve those algorithms. These ideas are essential to understand the objective of this proposal and research.

2.1 Research Objectives

2.1.1 Evaluate Algorithm Auditor Methodologies

Like any profession, the auditor has a series of tools and practices when conducting their audit. These series of tools and practices follow methodologies that are either documented or at time might be intuitive. These tools can include interviews and artifact gathering with developers, evaluation of systems using data, or red teaming as efforts to expose vulnerabilities with the system. As I aim to improve the practices of the profession, the first step has to be **understanding algorithm auditor needs**. Identifying needs justifies exploration into relevant areas from a practitioner's perspective. It also motivates and creates a philosophy through which a tool's ethos can be centered around. It serves no purpose to create tools to an audience to which they will not utilize. In order for Algorithm Auditors to have appropriate tools made for them, user needs and information needs to be gathered, assessed, and composed. This leads to several relevant questions:

1. What are the current practices for algorithm auditors?
2. Can those practices be concentrated into established methodologies and concepts?
3. What individuals beside the auditor are necessary for an audit to be successful?
4. How do current algorithm auditor practices relate to academic literature on the practice?

2.1.2 Evaluate LLMs on the basis of traits that leak information

In order for a system or technology to be considered equitable, it needs to satisfy a few constraints. One of those constraints is access. Access includes many elements such as the availability for usage. If for instance a technology is region-locked or requires intensive resources on the user end to utilize, the technology is not readily accessible to large swathes of the population. If it is marked that this technology ought to be a good that people should or need to have access to, then availability becomes a necessity. In the case of LLMs, since they are available locally for those who have the compute power, with significant work being done to concentrate on low-resource environments, and accessible for virtually everyone who has an internet connection, it is very much available. A second aspect to equity with regards to accessibility is 'usability.' These are reflected in making

¹The Responsible AI Suite can be found here: [Link](#)

²This Secure AI Framework includes a Map here: [Link](#)

sure various groups are able to utilize the technology with the same amount of fidelity compared to the base user. These groups could be individuals with different disabilities or those who utilize it through different modalities, such as speech recognition as opposed to typing as an example. Another group to consider are user groups that might have slight differences due to localities. When applied to technology like LLMs, this looks akin to user who interact with the model utilizing their language, whether it be a separate distinct language, or a dialect of a standard language. An aspect to be aware of is that due to the input being different, it may leak information about the type of user who is making the query, thus possibly biasing the output as well. This can inhibit equitable access since it can produce separate experiences for users on this basis. Equitable access is more than the availability of technology, but whether that technology is usable for people. In order to actualize this idea of **equitable access**, it must first be assessed as to whether these systems currently have differences in output for these users. This leads us to a few key questions:

1. Where can different user groups emerge when interacting with LLMs?
2. Is there a way to measure or approximate parity between these inputs?
3. Do those different user groups face parity when interacting with these LLMs?

2.1.3 Create Tools to evaluate LLMs on relative performance

After identifying methodologies and needs as listed under the first research goal, it then becomes prudent to then craft tools that coincide with that philosophy and methodology. This then means using the methodology extracted from the second research goal and making the process replicable. This target of relative performance is not to make a model aim to train itself to bring other categories of dialects to the lowest performing category, but to raise the less accurate categories in order to improve the model across user groups. To make algorithm auditing a more frequent phenomenon, tools needs to be created to identify the issues many auditors tackle and perform those tasks with efficiency. This is how we can proactively address equitable access to algorithm-based tools in the future. This will of course have the effect of improving accuracy of these tools across user groups, but makes these models more robust in general.

2.2 Expected Contributions

Contributions from this proposal will include the creation of a dataset, the creation of the benchmarking tool, and a clear signal towards a new avenue to which auditors can explore potential biases of LLMs in particular. The dataset in question is meant to host parallel prompts that are functionally equivalent. These functionally equivalent prompts can be used for training models and improving the presence of dialect specific words that are rare or underrepresented. Creation of a benchmark adds an additional tool that auditors can utilize to measure and understand the performance of a model. With a specific interest in dialects, it both raises the awareness of this problem existing and limiting the capacities of Large Language Models, while also helping auditors better identify those models. By having multiple categories being evaluated with this benchmark, it also signals which aspect of the model needs additional attention relative to other tasks. These pieces signal towards increased consideration toward these user groups, ideally making it easier for developers to be conscious of these groups at the outset of training as opposed to efforts of remediation after the fact. This is important because we believe that access to computing and technologies is a growing necessity for communities, thus expanding access has a multitude of benefits, including improving productivity, and lowering the barriers between users and meaningful interaction with models. Not only does this improvement expand the number of users who can utilize these models with greater accuracy, it has the potential to improve the experience for all users, since it provides the ability for models to increase their robustness against perturbations.

3 LITERATURE REVIEW

3.1 Algorithm Auditing

Algorithm audits are defined by Brown et al. as *"assessments of the algorithm's negative impact on the rights and interests of stakeholders, with a corresponding identification of situations and/or features of the algorithm that give rise to these negative impacts."* [67] It then follows that auditors are the practitioners who are conducting these assessments to spot bias and potential harmful effects of models. The literature surrounding algorithm auditors is centered around several areas: Case Studies, AI Policy, concepts & frameworks, and software. Each contribute to the thread of understanding the largely academic perspective of the profession.

Case study papers follow researchers attempting to audit some database or algorithm in some way, with their main contribution being the results from that audit and discussion of what the results could mean for the actors involved or society as a whole. The purpose of case studies are meant to explain and test a methodology that was utilized to examine the efficacy of these algorithms across certain groups through utilizing similar but distinct queries or input information on a machine learning system. Often these case studies are focused on a particular model for inspection, but could be broadened to look at a whole category of machine learning systems across different providers for the basis of comparison. An example of this is Godinot et al. [24] which takes time to highlight the risk of ML models and attempts to audit several different types of web-based machine models. These case studies highlight discrepancies between the types of inputs processed and spends much of its time within the discussion to elaborate and understand why the observed discrepancies took place. Typical reasons that are explained are extensions of machine learning principles in general. Phenomenons like dataset bias, where a lack of training data for a particular group or classification leads to the model having a tendency to misidentify the class in some way. Case studies in this context attempt to highlight these differences and offer input towards how this issue originated or recommends mitigation methods for future implementations.

Since a key motivating factor for conducting audits is related to compliance to local laws, AI policy is an important group of papers outlined that deal with auditors and standards to adhere to when administrating audits. Policy describes rules and practices that bodies have to abide by. These rules can come from organizations whose participation requires following, but the relevant examples we will be discussing are policies deployed by governments from various nations and political entities such as municipalities, states, and cities. Being in violation of these rules can come with consequences such as fines or the body being unable to operate within that nation or state. Due to the emerging policy from the European Union, Canada, and other governing bodies regarding the responsible use of AI, several papers concerned themselves with algorithm auditing as it pertains to how they might be utilized in adherence to these laws. Hartmann et al.'s paper [31] is a specific example of this, as it attempts to combine the frameworks of audits with the specific verbiage of policy to gain a greater glimpse at what specific actions auditors should do when attempting to uphold regulation as it concerns the EU's AI Act [19]. From the US perspective, [27] is an exploratory set of interviews that looks at algorithm auditors' impressions of proposed regulation. This paper has specific interest in laws concerning the hiring of individuals utilizing machine learning in some key capacity, which again crosses into the intersection of machine learning and Applicant Tracking Systems. This comes in response to New York City's Local Law 144 on Automated Employment Decision Tools (AEDTs).³ Since the US does not currently have comprehensive law regulating AI at a federal level, the paper provides a unique space to contrast to other Western countries and their respective approach.

Before a plan to conduct an audit is constructed, it becomes important to first lay out an approach to the process. This is a task both philosophical as well as literal as the task is informed both by constructing motivations of what the goal of the audit is as well as a process to actually achieve that goal. This is why papers concerning the concept of an audit, and their subsequent framework become important to analyze. Concept papers come without specific frameworks or software specified, thus can relate more to the concept of what algorithm audits should be and what they should accomplish. Rai [62] is an example of the general call to action to enact auditing as a procedure. The paper highlights particularly unique problems concerning AI and the lease of autonomy it is given for decisions, how responsibilities could be allocated when issues arise, and broad suggestions on how to execute an ethical audit of AI. Raji et al. [63] takes steps further by designing an ecosystem that can best accommodate audits featuring both application fields of audits as well as specifications of the audit type being conducted by auditors. This paper provides precise information for systems that can enable audits to take place, but are still differentiated from later categorizations as it does not provide specific actionable steps. These papers serve in some way as foundational works through which other papers referenced heavily, either in their understanding in what an audit was, or utilizing the information from these papers as the basis to form their framework or enact their audit.

A framework, much like a recipe, is meant to be a document that provides the step by step instructions for performing a task. At a higher level, a framework can provide an approach to a problem e.g. Markert et al. [52], or specific instructions in solving the problem e.g. Rhea et al. [67]. Papers that are categorized as being primarily framework in purpose relate to algorithm auditing specifically by supplying a framework for how algorithm audits should be conducted. How high or low level

³<https://rules.cityofnewyork.us/rule/automated-employment-decision-tools-updated/>

these audits could be is dependent only on the author of the framework. Frameworks can provide instructions like tasking to interview users or developers e.g. LaBrie et al. [42], test models e.g. LaBrie et al. [42], Akula et al. [5], and provide reports or documentation to the company or party involved e.g. Markert et al. [52], Naja et al. [54]. Instructions could be more detailed, specifically cataloging information such as specific developers involved in features and developing a structure of accountability e.g. Markert et al. [52], Naja et al. [54], or running reports over databases e.g. Akula et al. [5]. All of these steps can be essential depending on how comprehensive the audit is meant to be.

The next step after establishing the framework is operationalizing the audit, which is often through not just the auditor conducting the actions outlined in the above frameworks, but being empowered to do such through the numerous software contributions that align with aiding auditors as well. These papers elaborate on technology that is created by the authors to assist in the auditing process in some key capacity. Their contribution is the software itself, with discussion on how it can improve the auditing process in some key capacity such as documentation, or in fairness assessments. With the smallest number of contributions, the software tools that help conduct audits are widely different. Naja et al.’s [55] works to aid auditors by providing a provenance based approach towards tracking and documenting the documentation auditors need to utilize when localizing possible biases in the ML implementation. This approach is informed by the necessity to document accountability for when instances of AI-related incidents occur, systems and points of contact can be established to know who is both responsible for implementing mitigation and for understanding who implemented the code. This provides knowledge of systems and projects beyond the minds of those who implanted these systems in a way more cogent than walls of documentation and blame commands within repositories. Hertweck et al. [33] and Salerio et al. [69] make more broad toolkits as their contributions to the field. These toolkits feature numerous perspectives in terms of users, varying from data scientists and ML implementers to policymakers who look to obtain perspective on the complexity of what are large technical systems. Algorithm auditors are one of the effective users of responsible AI technology. Their goals in auditing are to help identify issues within models, and as such are a key stakeholder as it regards development within this sector. It then becomes relevant to review both the work they have done and the research that has been covered, so as to find what is not currently being researched, and to see what highlighted problems pose significant risk for auditors or users.

3.2 NLP & Low Resource Languages

NLP describes the field that focuses on systems’ capacity to understand, interpret, and generate human language. This research has numerous fields such as machine translation for translating between languages, information extraction for tasks dealing with unstructured data, and spoken language processing, which represents another key modal form of interaction with models. These models can be analytical such as those that specialize in sentiment analysis [4], or they have the capacity to be generative like natural language generation [18]. What is especially important for this research are Large Language Models specifically, as they are the generative models that large swathes of people utilize. LLMs are systems trained on text to generate human-like language [1]. They read text inputs (prompts) as tokens and uses a neural network in order to predict the next token in its generation, which in turn generates text with the spontaneity that is similar to human-generated text. They are built on transformer architecture, which processes relationships among tokens in a sequence [75]. This emphasis on order helps provides context to understand which token the model should pay attention to, which is important. This then means that having enough exposure to those tokens becomes important because without enough training or exposure, then a model will be unable to recognize the token’s significance. This can be especially true in the context of low resource languages.

Low Resource Languages is a label to describe languages for which there are insufficient computational resources such as corpora, annotated training data, benchmarks, and language tools [49]. Examples of this can extend to many parts of the world such as Asia [7], Africa [36], and numerous Indigenous communities around the world [74]. It is especially notable of languages that do not have have a writing system attached to it, as this drastically reduces the amount of data that is historically available for analysis. The specific issue in NLP with regard to these languages is that since they lack the sufficient amount of training data, thus create greater issues for NLP applications to appropriately learn from. Solutions to address this are numerous such as utilizing synthetic data [45], cross-lingual approaches like transfer learning from larger models [3], and even utilizing communities and speakers from those communities to help create additional authentic data [12]. These efforts help make both language specific and general models better through the creation of high quality training data. This also creates a guide for other similar pursuits. An intersection where low language techniques becomes relevant for large standard languages are dialects. Dialects are defined as a variety of a language often associated with a particular region, social group, or community which can cause differences in vocabulary as well as grammar. Examples of dialects include African American Vernacular English(AAVE) [41], Appalachian English [32], Texas German [9], among others. Their origin is attributed to several reasons but often come as a result of the region, ethnic group, and laws or policies that might have dictated language use [34] [77]. While the line between language, sublanguage, and dialect can at times be hazy, it is important to note that these dialects can be written out typically (albeit often phonetically) utilizing the standard language’s typeset. Because of this, users who speak this dialect often interact with platforms in a way that mirrors those vocal spoken differences. While this was hard to document in the past, a growing body of evidence via social media displays the average user interacting on the internet using their dialect. This interaction is not

limited to between users however, but from users to models as well. This provides a justification for research into this specific intersection and use case.

NLP and Low Resource Languages have relevance to this specific field since the dialects I am looking at utilizing in DiaBench are primarily low resource languages. Research suggests that people utilize LLMs for tasks such as fact finding [15], among others. This creates an observable population who are interacting with large language models with dialects, and as such, provide justification for investigation into this research. This interaction is important because it then means that trust in these platforms is an important factor as users could be fed erroneous outputs at higher rates, thus signalling biased experiences for users.

3.3 LLM Benchmarks

Benchmarks are a standardized set of evaluation tasks and metrics used to evaluate model performance under certain conditions [1]. Benchmarks are important because they help users, developers, and customers with understanding how well the model performs relative to other models that have gone through the same benchmark. This is important as models that markedly improve on performance are verified in that performance once they display their competency on the benchmark. With regards to LLMs, benchmarks evaluate LLMs under conditions such as n-shot prompts, allowance to access external documents, time or cost of producing an output, and metrics standard to machine learning such as F1-score, as well as NLP specific metrics such as BERT [23].

Benchmarks evaluate LLMs on a number of tasks. Tasks include **language comprehension** [8] evaluating the model's ability to detect and fix errors with text, **code generation** and the model's ability to produce syntactically correct code that solves user requests [57], **Question and Answer** which looks to determine the model's ability to respond correctly to inquiry with objective answers, and **reasoning** which asks complex questions or riddles that have increased context or require logic or deduction to answer [30]. These tasks are relevant because they match the ways in which people utilize these models for, as various use cases of LLMs are driven around spell checking [87], Inquiry answering, generating code for novel or simple applications, or for explaining concepts through chain of reasoning. Each of these tasks are important and novel interactions with models that are beyond what is capable of simple queries to search engines. In their generative nature, LLMs can combine the context through which they receive as input with the information it might gather online to composite an original response in a way that justifies many user needs, thus LLMs are particularly important models to assess.

Metrics for evaluating these models are various depending on the task. If it is in the context of safety, a common metric is attack success rate, which looks at how often a model is unsuccessful in their ability to detect harmful speech or instruction [13]. Previously mentioned F1-score is typically used in classification tasks, and simple metrics like plain accuracy can be utilized for QA tasks. While using metrics help evaluate approximately how well models can perform on unseen information, they are not always reliable sources for actual model behavior. Issues with reliability stem from a feature in machine learning and reward functions. If a potential model is trained to pass benchmarks, this can reduce how well the model performs in the field, as the model will retrain itself to pass the benchmark rather than to be generalizable [84]. An issue tied to this phenomenon is contamination where models might incidentally gain access to the data from the benchmark, thus having seen it before testing, rendering the benchmark less effective. Current literature attempts to remediate this issue with privatization [64] or with constantly updating with real-world knowledge [79], but it is still a pertinent problem. This possibility for benchmark contamination and subsequent training makes it clear that benchmarks are a tool in evaluating a model, but are not the only tool that should be used for highlighting performance of a model. In the creation of a benchmark, seeing what other LLM Benchmarks do provides a guidance towards what this benchmark aims to do, both in terms of technology and in terms of assessment. Their perspective and aim reveal both the goal and intended audience to utilize these benchmarks. As they often are used by developers, it becomes a task to understand how auditors might utilize them in different ways to auditors with explicitly different goals in mind.

3.4 User Populations for Models versus Training Data and Auditing for Generative Models

The generative language models of today that have the largest amount of training data often are trained on sources that are predominantly of a particular lingual background, with over 90% of ChatGPT's training data coming from English content with a "US-Centric point of view" [17] as an example. Since many of these models are privately owned and often contain proprietary algorithms, we can only guess the sources, but since these models are trained from literature and the publicly available web, then it stands to reason that since the majority of the web is written in standard English, both American and British, then model training is informed by much of the written documents on the internet are also in the same manner. While this may be serviceable to users of those specific backgrounds, there is cause for a change in audience. As the internet and these models gain more users from a variety of backgrounds, this reflects a shift of user bases [26]. While this global community can be reflected in the multitudes of users on popular sites such as social media, this is not necessarily reflected in the training data for models that every user might interact with. This is because while these dialects and languages do not have a substantial history

of written documentation either physically or online. While a possible solution would be competing models that are trained with their own set of training data accustomed to the nuances of a particular language or dialect [11], larger models with more base knowledge should be adjusted to the particulars of the users rather than models being created from the ground up since this increases the general usability of the model. This means to say that in order to improve the experience in users and their ability to utilize these generative models, then it becomes the charge to improve these large systems with vast training data. One of the ways improvements can be delivered can be through audits of these systems.

Generative models represent a new paradigm for users’ interactions with artificial intelligence and machine learning systems. The ease to which users can interact with these systems have in kind increased usability to the point of public awareness and subsequent saturation. Their performance can fall to certain behaviors however. Inputs both intentionally erroneous as well as coincidentally erroneous, can produce hallucinations [86]. One of the key drivers that can help users utilize these models is their ability to interact with the system using systems they already know. In the instance of the LLM, the system users already know is language, and that becomes their mode of interaction with the model. However, not all representations of a language are as well-covered as others within a system’s own training data. This level of comprehension is inhibited when the input language has slight variations to the language, either by error, such as typos [22], or by design as is present with dialects. As a result, prompts that mirror dialects as input to these models can reduce the model’s ability to comprehend the prompt and be more likely to produce hallucinations. This erodes user trust, thus reducing how often these tools are likely to be used. In order to both increase trust and usage across these user bases, we introduce the following studies to first identify, document, and help remediate issues with LLM performance. This manifests through our interviews with algorithm auditors, utilizing those findings to inform a methodology to document issues with LLMs with regard to this particular issue, then attempt to democratize this process through the release of a benchmark.

4 Proposed Work: Interview Study on Algorithm Auditors

Gaining a perspective of algorithm auditing is more than an inspection at academia. Since it exists at the intersection of businesses, governance, policy, and technology, then it becomes pertinent that the perspective of practitioners should be collected. In order to capture this information, several interviews of individuals related to algorithm auditing, AI model development, and AI Governance were gathered. This core of professionals provide a diverse set of experiences that enrich the data with nuance of what auditors do in their evaluations. With 9 professionals interviewed, we have collected a body of information that we can analyze. Current work is primarily being performed on coding the responses from the interviews, and extracting themes that appear. This investigation comes with a few key research questions:

1. **What is the current documented process for how algorithm audits are employed in practice?**
2. **What are the resources utilized by algorithmic auditors in determining their framework aka the regularized process through which they audit an organization?**
3. **What importance do subject matter experts, people who are experts in the field that pertains to the specificity of the data, play in the auditing process?**
4. **What is the technical background/makeup of those who conduct algorithm audits?**

What we aim to obtain from these interviews is an inspection of what tasks auditors currently perform, their current challenges, observations of compliance, and the needs they want met in order to better audit systems. These observations can help us adjust the user requirements for benchmarking tools as well as assess other considerations for improvements in algorithm audit design and implementation.

4.1 Background

Algorithm auditors as professionals have a wide range of experiences as it relates to conducting audits. Techniques in auditing are informed by other auditing practices such as accounting techniques as well as software testing. These backgrounds demonstrate how the auditing process integrates both the machine learning specifics of model assessment and analysis with forensic approaches commonly associated with financial accounting. While the exact practices employed in an audit vary due to the type of model, application, and relevant policies that exist, there are several shared elements. One of those shared elements are the expectation of a report and recommended actions to remediate. What this looks like in practice is a report on notable metrics such as disparate impact and if the model outputs exhibits an unfair bias against persons or certain groups, or if it contains biases that limit its functionality. These audits test models in numerous ways in order to validate each of the processes involved with building, training, and releasing a model for use. Training datasets can be inspected for integrity by performing data audits to ensure data is homogeneous in shape, high quality, representative, and complies with policies concerning Personally

Identifiable Information (PII). Due to the black box nature of modern models, it is difficult to analyze layers in a transparent manner. Thus, auditors look to the elements that can be viewed and are human interpretable: the training data, the input, and the output. Models can be inspected by their outputs given directed input. These outputs are used as the basis of comparison to approximate different experiences for users to detect anomalies that can be correlated to attributes such as gender. If given access, models can have their hyperparameters inspected to see their influence on the outputs generated by the model. Although they do not provide a direct relation to metrics due to the black-box nature of these parameters being tools for fine-tuning, it can help provide a picture into the range of possible outputs. Entire systems can be checked for their potential security vulnerabilities to ensure malicious actors cannot poison training data, take advantage of the model, or deploy versions of models not meant for the public. To that end, an audit encompasses more than what would be expected of debugging code and can be seen as an end-to-end inspection of a model at their most low-level implementations. In addition to these practices which involve looking at the technical requirements of the system, there is also an analysis of the surrounding structures that make up the system. Taking account of who in the organization is responsible when these systems produce faulty outputs becomes an important measure [51]. In addition, understanding the needs of the various stakeholders such as users as well as developers are measures auditors take account for when conducting audits [43]. All of these processes are meant to help increase both developer and user confidence in these machine learning models as they provide recommendations or decisions in critical applications.

Current literature surrounding algorithm auditing from the perspective of understanding auditors is limited. Groves et al. [27] conducts interviews of algorithm auditing professionals, but looks to their opinions and thoughts on policy concerning New York’s local law which regulates usage of AI in the employment process. While these issues are important to understand and study, it does not cover the complete range of algorithm auditing that is currently being conducted. In addition to audits being conducted to keep within compliance to law, audits can be done voluntarily, thus analysis of auditors who perform those audits becomes prudent. Research related to auditing conducted by companies within the corporate setting provide practical guidelines for practitioners to abide by such as Ashar et al.’s contribution [6], but these guides are recommendations of practices, not an analysis of what current practices are being administered. The reason for this limited amount of research is in part due to the recency of the development of the practice. This paper’s research is necessary as few papers look to interview professionals in an effort to document and understand the techniques utilized, and to bridge the gap where current algorithmic auditing literature exists. In understanding that gap, we can evaluate and understand what algorithm auditors need, and what academics and literature might be insufficiently covering.

4.2 Methodology

4.2.1 Participants

Algorithm auditors exist as a relatively new profession. Because of this, the search for appropriate professionals was extended to those involved with the AI risk management process at any level. Searches for participants took place using several methods. Investigation of notable algorithm auditing firms and contact with the organization was one of those key factors that helped provide participants. Searching through other professional networks involved in responsible AI, as well as through directly contacting professionals on networking sites such as LinkedIn. Due to this expanded search a multitude of specialists were interviewed in this study.

Participants in the study have a range of experiences, with backgrounds in algorithm auditing, AI risk management, as well as AI red-teaming. The application spaces of these systems vary with areas such as healthcare technology, social media recommendation, and cosmetics in addition to audits done at the behest of relevant policy affecting the system. These backgrounds give an insight to understanding the wide array of applications as well as different parts of the auditing process coming from different roles that contribute to the overall audit procedure. A total of 9 participants were interviewed over the course of this study.

4.2.2 Interview Methodology

The interviews take place one-on-one between one of the study administrators and the participant. The interviews each take approximately an hour to complete. Since the structure of the interview is semi-structured, the questions can be adjusted to the interviewer’s background. Participants are asked for consent in recording before the interview can begin. Recordings in addition to transcripts assist in the process analyzing answers. Questions belong to these four key areas: Background, Audit case study questions, Audit design, and post-audit process. Notes are recorded by the study administrator during the interview for initial collection, but transcriptions of the interview are also created to fill in gaps in the answers that are recorded. After interviews are conducted, answers are analyzed by the study administrator for thematic synthesis. This is done to find common themes among the interviews. To view the entire list of questions and interview procedure, refer to Figure 6 in the Appendix.

4.2.3 Question Design and Goals

Questions for the interview are meant to inspect both the background of the auditor as well as gain insight into each aspect of the audit, from its conceptual ideation to its specific application and subsequent off-boarding process when the audit is finished. Questions surrounding **background** are asked to get an understanding of the interviewee's experience, be it with machine learning, with policy regard ai systems, or with specific subject matter. These questions help extract the most information from individuals in the following sections, as some technical questions such as technical processes to test a model would not be as appropriate for individuals whose responsibility is primarily fielding user sentiment as an example. **Audit case study** questions narrow down to lower level of the audit process. Interviewees with an extensive background in auditing are asked to talk about one of the audits they have conducted with some caveats. No information concerning what company, beyond its sector (e.g. healthcare, hi-tech, finance, etc.), is being audited or anything that might be considered proprietary to the auditor's process. These questions are asked to get an understanding of which broad applications of machine learning are being utilized and the types of specific challenges both auditors and firms face when attempting the audit process. After asking about these specific scenarios, we zoom out and inquiry about **audit design**. Interviewees are asked in a bit more explicit detail the process of how they conduct an audit. This includes information such as the features of their audit, what frameworks they utilize, as well as the philosophy behind their auditing procedures. This is done to find commonalities between auditors as far as specific techniques and approaches are concerned. To finish the interview, we focus on the **post-audit process**. Interviewees are asked about what interactions with clients and models after the initial audit process. This process is mainly off-hands, as clients deal with the results of their audit without necessarily needing to consult with the auditor further. It is still important to garner information from this stage of the auditing process because it is used to track which barriers firms faced with implementing changes suggested by auditors.

4.2.4 Analysis

After intake of questions and answers from both notes and transcripts, there are two layers of analysis we conduct in order to extract the most information from these interviews. The first of these analyses are a question and answer coding. We specifically look at the questions originally asked in the interview and pair them with the answers obtained from the transcripts, pulling codes from those answers. This coding helps us find overarching themes commonly found among interviews in order to help us answer our stated research questions. The second layer of analysis is more granular, as we utilize RTA (Reflexive Thematic Analysis) [10] in order to search transcripts line by line to find codes that were not explicitly part of the question and answer process. It is important to conduct both since it provides both overarching information in addition to novel information that will be important to note.

4.3 Results

4.3.1 Results: Interview Profiles

Table 1 highlights information such as industry experience as well as background. This is done since they interact with the themes of the paper. One of those themes being the use of **interdisciplinary work**, varying both with backgrounds in technical fields as well as traditionally not associated with computing, as well as applications. These multitude of experiences provide further proof of the vast number of applications machine learning has ingrained itself into to such an extent, that it needs auditing.

4.3.2 Results: Interview Codes

Questions have been categorized according to their area listed created in the initial interview document. Each of these sections are a code for the type of question asked to the interviewee and guide the codes we create for the responses during the interview. Typically, these codes contribute to an overall theme of an answer rather than specific information from the answer. For example, a notable code that has emerged in several questions and answers has been **variability**. Since auditors conduct their practices with real companies, the way those companies have implemented their model, have created their documentation, as well as how they have stored important information are pieces of a puzzle that change with every client auditors are introduced to. Codes like these are extracted from the answers and grouped to provide a coherent picture of what the industry at large is facing. While codes are still being extracted, many of them are included in this work. Table 2 shows a sample of both these questions and the associated codes that emerged from the interviews. In total, there were over 100 unique codes extracted via both answers analysis and through the RTA methodology. While some were only mentioned in one interview, they were provided significant context towards the greater themes surrounding these interviews.

Participant	Title	Background	Audits Conducted
1	Product Officer	Technical	Finance, Hi-Tech, HR
2	AI Risk Assessor	Non-Technical	Medical AI Applications
3	Data Consultant	Technical	Social Media
4	Strategist	Non-Technical	CPG, Social Media
5	RAI Engineer	Technical	Government, Energy
6	Tester	Non-Technical	N/A (Participant)
7	Software Engineer	Technical	Networking Systems
8	Auditing Company Founder	Technical	Finance, Hi-Tech
9	Professor & Anthropologist	Non-Technical	Social Media, Moderation

Table 1: Interviewee profiles and notable audits conducted

Question	Question Category	Associated Codes
Do you have a lot of experience conducting audits?	Background	Technical (Programming) expertise, Interdisciplinary experience
How long are engagements for audits?	Audit Case Study	Variability
What were some metrics that you needed to collect?	Audit Design	Flaws in data collection, Stakeholder-based
Where do firms often fail in terms of their audit?	Post-Audit	Technical Documents, Wide scope

Table 2: Examples of questions that were asked to participants concerning conducted audits

4.4 Discussion

Several themes emerged from collecting information laden within the interviews. While not every theme will be discussed, the most pertinent ones that we discuss create an opportunity to see exactly what auditing looks like, the challenges it faces, and the philosophy guiding the frameworks and methodologies of these auditors.

4.4.1 No Set Standards

Something observed through the interviews was that beside audits that are done in direct response to policy, with stated metrics to calculate for firms to aspire to, there was no set standard for how audits should operate. This is due to both the abstract nature of audits, as well as ill-defined audit standards. Between all of the Participants, none shared a single standard or framework with another. While some publicly available standards were mentioned, no two auditors shared the same standard. Instead, auditors often made their own frameworks and processes to conduct an audit, which itself was subject to a great deal of variability. Participant 05 made a statement as such stating, *"there's not really a one size fits all metric to look at the bias of a given algorithmic system. The whole point of that process ... is so that we find the right metrics for this system in this context."* This means that there are currently few standards universally applicable when an audit is conducted. This is despite the fact that there exist many ways to evaluate a model's performance from a technical standpoint as made aware through benchmarks and other technical aspects. This observation is made in part that because the process involving more than just models, but rather people who create models and applications that can be novel, there are more concepts to share rather than set standards or instructions. It remains a point to inspect however, as this naturally means it is thus hard to compare one audit conducted by one auditor to another's.

4.4.2 Stakeholder Involvement

A notable code that became frequently referenced was **stakeholders**. Stakeholders were either explicitly mentioned in interviews or alluded to in terms of procedures and priorities that had to be accounted for. Stakeholders refer to people who are affected by the model in some articulated way. These stakeholders are usually split into two areas: Internal and External. Internal typically refer to those internal to the firm such as a company's legal department or the engineers creating the model. Participant 05 talks about them in the context of a hiring model, saying, *"Let's just imagine a company is using an algorithm to do some HR and hiring stuff...There's probably internal stakeholders like the HR Department ...They want it to work, you know. They[HR] want to hire high quality candidates."* External stakeholders would qualify as people who are subject to the model or exist outside the organization. Participant 05 elaborates stating, *"...external stakeholders like the people who are actually applying to the jobs at this company. They're concerned whether the AI hiring tools are working properly."* Their involvement is essential because they provide the perspective of both expectations as well as fears people have about these systems, and can

exist as the seeds through which an auditor can validate and explore an issue. Participant 14 affirms this idea as well stating, *"You should always start with identifying the stakeholders...at least half of the risk assessment process is engaging with the stakeholders and collecting their views."*

4.4.3 Subject Matter Experts

Since many machine learning models are deployed in applications that require knowledge of the subject area in addition to knowledge of the model, it requires a unique set of information involving both. Since algorithmic auditors are primarily knowledgeable about models or fairness, a notable observation was how many auditors either requested or brought along subject matter experts to aid in the auditing process. Some were working in tandem with the auditor within the firm, or brought externally. Participant 08 describes the necessity of their inclusion saying, *"...there are some areas that you know we're not [sic] we're not gonna be comfortable with making opinions, you know, making judgment calls about. We're gonna need... subject matter expertise to work with others to then help them form, you know better inform decisions and judgment calls and opinions."* These subject matter experts then become a backbone for making sure an audit is successful in achieving its goal in validating and attempting to remediate concerns. This once again demonstrates how relevant interdisciplinary work is when conducting algorithm audits.

4.4.4 Context Sensitivity

Application and context were frequently occurring themes in interviews as well. It mattered not just the actual model's capabilities, but the purpose for which those who had access to the model were going to utilize it. Participant 05 articulates about this about a potential health metric, *"...it has a score [of] the chance that that [sic] person is going to develop diabetes in the next couple of years...Let's say that we were asked to audit that...you have to specify a context because If that score is in the hands of your doctor then that could actually help you... But if that very same score is in the hands of a potential employer who's deciding whether you're gonna be like a costly insurance policy for the company. They might use that very same score to not hire you, and that would be bad for you."* This idea of context extends past the idea of application, but to the metrics utilized as well, as Participant 08 says *"[disparate impact] is not the best sort of like the most sort of like appropriate metric. Considering the used case, for example, of a certain algorithm, or you know, because of certain output that they get you don't. You can't really, you know, use this for impact."* What becomes clear is that an audit involves understanding the intent of the application and provides a delineation that certain models and data are usable, but should prioritize the safety of persons subject to it.

4.4.5 Security

An important avenue of algorithm audits that incorporates information from AI risk management, would be assurances and protection against adversarial action. These adversarial actions can be in the form of input to purposely illicit irresponsible responses, or the generation of output that can be considered in breach of policy or toxic. Participant 22 spoke candidly on the subject, saying that *"AI systems are often given too much privilege,"* detailing how individuals provided too much in terms of permissions to data, which, if used by some adversary, could access both identifying data as well as sensitive attributes. This angle was novel among the interviews, but is a necessary point of discussion since it involves another perspective that builds support to the idea of an audit as a holistic inspection of not just data, models, or even individuals, but policies and practices as well.

4.4.6 Top-down vs Bottom-up: A Philosophical Approach

A division between those who had technical backgrounds and non-technical backgrounds was the process of approach when conducting an audit. Technical background in this case refers to whether the individuals conducting the audit had machine learning and programming expertise prior to working within the audit space. Auditors with a technical background often discussed the process of auditing, and their subsequent for auditing based around a bottom-up approach, that being the process was concerned first and foremost with the testing of systems and capabilities, while auditors from other practices who have come to this profession exhibited a top-down philosophy.

A top-down philosophy seemed to focus on the organization first, the parties involved with the creation of the model, and the involved stakeholders. Stakeholders refer to parties that have some interests relevant to the model, such as the team responsible for its creation, managers, or users of the model. This observation spoke to the fact that unlike other practices such as model debugging or security testing, algorithm auditing involved more interdisciplinary participation since it involved both creators, users, and sometimes customers, all whom have varying levels of technical expertise and different needs. This marked difference should leave to discussion that with these approaches, goals start to diverge as well, since bottom-up approaches were more concerned with metrics of the model while top-down approaches were concerned with inspecting institutional or systemic problems that could exacerbate the consequences of bad model outputs.

4.4.7 Policy-Motivated Audit vs. Voluntary Audit

Another division that emerged when interviewing auditors were the specific customers these auditors were servicing. The majority of those who approached for audits were primarily concerned with regulatory compliance, as model deployment or usage in many instances would need some auditing measure done either before or shortly after deployment. These customers, although complying to regulation, expressed more apprehension to major changes to systems and initially saw less benefit to the suggested changes by auditors. Justifications varied, but a consistent observation from those customers were centered around cost and resources required to remediate the issues being too high.

This contrasts to audits that were conducted voluntarily by the firm. Audits that were in this area included inspections for AI-guided CPG product lines or with other AI-enabled technologies that did not have regulation attached to their implementation. This curated a different experience for auditors and for firms, since it was a process for understanding a specified goal or task that was originated by the firm. The suggested remediations and results for those audits also varied. While one can assume that those voluntarily requesting an audit are more prepared and likely to implement advice after the conducted audit, the extent to which this adoption occurred still varied due to similar reasons relating to cost to remediate. In some cases, the motivation behind the audit was not to improve these systems, but rather as an attempt to signal their forward thinking nature by conducting an audit in the first place, which displays a motivation driven more by reputation than by performance.

4.5 Conclusion

Algorithmic auditing as a practice is new enough that the current practitioners are fashioning methods as they go along. These methods they create are informed by previous experiences in research, accounting, and model building. While each has their own base to build on, the lack of standards make audits hard to compare to each other, even if they share model performance metrics. In order to gain a full assessment of a system, it requires the work of more than just the auditor themselves, but also the knowledge of subject matter experts who know the application in context in addition to stakeholders. These parties are all necessary collaborators from making a system that functions into a system that is both successful and responsible. With fairness and bias at the forefront of their concerns when conducting audits, practices and motivations are centered by outcome and real application rather than abstract possibilities.

5 Proposed Work: Testing LLMs on QA in English Dialects

With expanding user groups interacting with nascent technology as internet access increases, it becomes important to understand both their user habits as well as how systems interact with those habits. For instance, if we look at dialects and languages presented by user groups such as West Africans and West African communities abroad, interactions online show usage of dialects as a mode of communication with each other and with systems. This investigation of LLMs on QA in English dialects looks at assessing comprehension of dialects through the accuracy of QA responses. The questions are retrieved from a known QA dataset, and are translated into the target dialects through a combination of both NLP models as well as local speakers. Once translated, they are sent into several popular publicly used LLMs to net a response. These responses are evaluated both on the accuracy of the prompt response (correct or incorrect), whether the response matches to the dialect of the input prompt, and in the case that the question is answered incorrectly, the type of error that can be attributed to the incorrect response. An earlier iteration of this experiment has been completed [16] with an additional study created to both increase the scope of the questions as well as to expand evaluation of the response. The earlier study was done in collaboration with other researchers at the institution, with experts with more domain knowledge in Natural Language Processing (NLP) being involved. My specific contribution towards the experiment was the research lead, formation of the methodology, sourcing the dataset, script writing for prompting, evaluating a selected number of responses, analysis processing, and writing a great deal of the paper. Collaborators contributed through writing, research concerning dialects, as well with creation of figures displaying the methodological flow, which are included below.

5.1 Background & Related Work

This study examines discrepancies in LLM performance with AAVE and WAPE, building on emerging work that probes the consistency and coverage of LLMs in non-standard dialects of English. Prior research informs our approach by demonstrating how others evaluate LLMs on different dialects, guiding how we think about achieving more consistent behavior across a broader range of inputs.

5.1.1 Linguistic Background

AAVE and WAPE are largely absent within LLM training data. About 30 million African Americans use AAVE in the United States [78]. AAVE originates in the American South, where enslaved Africans learned English vocally. This leads to a spoken form of English that blends with Southern speech patterns. As a result, AAVE has distinctive vocabulary, grammatical structures and punctuation patterns.

WAPE, also rooted in oral traditions, is spoken by an estimated 140 million people across West African nations and in African immigrant communities originating in vocal communication [83]. WAPE often features phonetic spelling, such as “them” becoming “dem,” and typically omits definite and indefinite articles unless emphasis is required [21]. These grammatical and lexical differences make AAVE and WAPE more likely to be misinterpreted by LLMs not exposed to them during training. AAVE is a culturally rooted dialect with distinct grammatical and phonological patterns that differ from SAE, posing unique challenges for language models not trained in its structure. These may not seem like significant changes, but in Standard American English, grammar plays a large part in the meaning of a sentence, which can alter the way an LLM interprets it. Some common grammatical differences include the use of the *Coulpa BE* Absence, double negatives, and a general negator [70]. In AAVE, the verb “to be” is often omitted when referring to the present state. This is known as the *Coulpa BE* Absence. For example, in SAE, one may say, “She is my sister,” whereas in AAVE, it may be altered to “She my sister.” Another common rule of AAVE includes double negatives, which is the use of multiple negatives for emphasis. This is seen as grammatically correct in this dialect. While these are some overly simplified rules, there are some occurrences that are a lot more nuanced and have the ability to change the way an LLM may interpret input made by someone using AAVE.

Similarly, WAPE is spoken by about 140 million people in West African nations and African immigrants originating from that region [83]. It serves as a *lingua franca*, a commonly adopted language between non-native speakers, for communities that contain hundreds of individualized languages [83]. It has origins in commerce as West African tribes communicated with British traders, and like AAVE, evolved without formal written tradition [83]. As with AAVE, WAPE has its own vocabulary and grammatical rules. The most prominent difference between SAE and WAPE is the spelling. Because WAPE is essentially a spoken language, words are often spelled based on how they sound [21]. For example, in SAE, words like “them,” “my,” and “we” would be spelled like “dem,” “Mai,” and “mi” in WAPE. Another common grammatical difference between the two is that definite and indefinite articles (a, an, the) are often omitted in WAPE unless emphasis is needed [21]. For example, in SAE, one may say, “I saw a dog in the street,” whereas in WAPE, they would say, “I see dog in street.”

5.1.2 Related Work

Gupta et al. [29] develop the AAVE Natural Language Understanding Evaluation (AAVENUE) to assess performance on natural language understanding tasks in AAVE. Our study builds on this work, examining whether performance gaps persist across dialects in a QA setting. Lin et al. [46] examine LLM performance on tasks like logical reasoning and math when prompts are written in AAVE. They compare model performance demonstrating noticeable drops with small differences in how the prompts are written. These findings guide our decision to directly compare the LLM outputs across dialects rather than relying on prompt translation.

Research on WAPE usage in LLMs remains limited, with newer publications and pre-prints identified. Adelani et al. [2] develop a benchmark translating SAE to WAPE and Naija (another common Nigerian language) and test whether the WAPE-trained models also perform well for text generation. Our study takes a related approach by testing how dialects within a shared diaspora—with uneven training coverage—affect model responses. Lin (Pin-Jie) et al. [47] show that models tuned on Nigerian Pidgin outperform multilingual ones on dialect-specific tasks. Few studies compare LLM responses to equivalent prompts across SAE, AAVE, and WAPE. Additionally, no prior work has directly compared LLM responses to the same prompts expressed in SAE, AAVE, and WAPE side by side.

5.2 Methodology

Dialects express similar intents in ways LLMs may interpret inconsistently. We simulate this variation by creating equivalent prompts (see Figure 1) and comparing performance against SAE, the baseline given its dominance in training data.

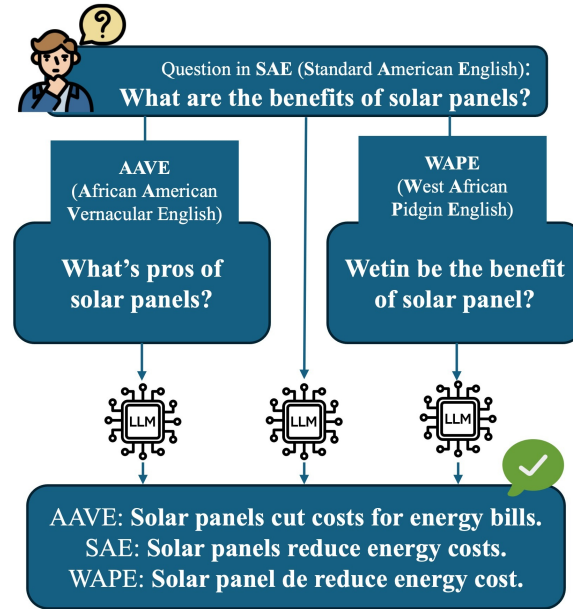


Figure 1: Dialect Translation and LLM prompting

Assuming AAVE and WAPE yield lower QA performance than SAE, this study frames the following research questions: *To what extent do LLMs exhibit lower QA accuracy on (a) AAVE (RQ_{AAVE}) and (b) WAPE (RQ_{WAPE}), in comparison to LLM performance on equivalent SAE prompts?*

5.2.1 Study Design

This study aims to measure accuracy across different LLMs through a sequence of steps: selecting appropriate LLMs and a dataset to translate, applying dialect translation, and conducting evaluation (see Figure 2).

5.2.2 LLM Selection

We select LLMs based on their popularity and general accessibility to the public: ChatGPT, Grok, and Gemini. These more closely reflect current public versions of LLMs compared to enterprise models.⁴

⁴ChatGPT (OpenAI), Grok (xAI), and Gemini (Google) are accessed via public APIs under their respective non-commercial research terms and acceptable use policies.

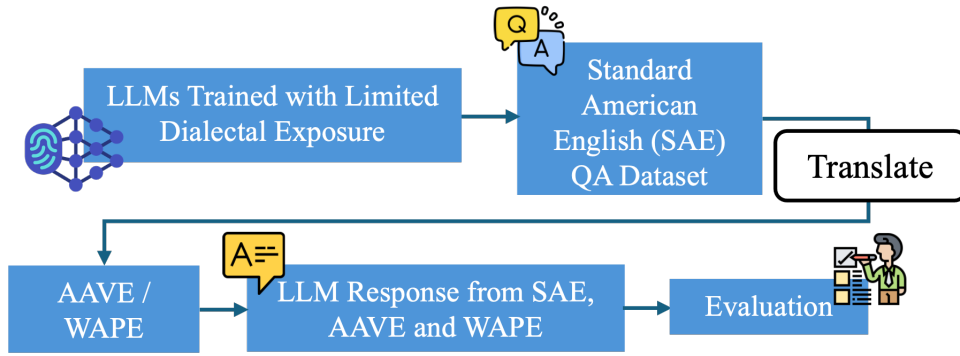


Figure 2: Methodology Flowchart

5.2.3 Data Collection

We select the Stanford Question Dataset (SQuAD) [65] for its wide use in LLM benchmarking, objectively verifiable answers, and suitability for consistent human evaluation. Its fact-based questions allow clear accuracy judgments and help identify hallucinated outputs. With this dataset, we translate queries into AAVE and WAPE for further analysis.

5.2.4 Question Dialect Translation

We translate 1,000 randomly selected questions, using their original form as SAE prompts. AAVE versions are generated with an online dialect tool,⁵ and WAPE equivalents are produced via the PidginUNMT project,⁶ shown to yield high accuracy [56]. We process each SAE question using these tools to produce content-equivalent prompts with full examples found in Table 3. A speaker of these dialects then reviews the post-translation questions to validate. These speakers are not linguists, but casual speakers who know the dialect, so the process involves evaluating if the translated prompt “sounds” roughly correct. Once validated, each question is sent to the LLM via an API, with a system statement constraining responses to fewer than 20 words. This prompt is written in SAE across all dialect groups to simplify evaluation and ensure consistent comparison of short ground-truth answers with longer LLM outputs. The prompt is queried without the surrounding context featured in SQuAD since additional context creates additional ambiguity for the dialectal equivalent phrases. Each response is recorded and compared to ground truth for evaluation.

Dialect	Question
SAE	Along with drought, what is one other factor that is pushing the Amazon rainforest towards a tipping point?
AAVE	Along with drought, what else pushin’ the Amazon rainforest toward a tipping point?
WAPE	Along wit drought, wetin na one other factor dat na pushin’ di amazon rainforest towards a tippin’ point?

Table 3: Example SAE question with WAPE and AAVE equivalents

The resulting dataset consists of 1,000 SAE questions translated into AAVE and WAPE. For this paper, we evaluate a random sample of 100 SAE questions and their dialectal variants, yielding 300 prompts in total. With three human raters scoring 900 responses, this represents the largest feasible scope given our resource constraints.

⁵Clickable link: [Mr.Dialect](#)

⁶Clickable link: [PidginUNMT](#)

5.2.5 Evaluation

We evaluate 300 prompts against ground-truth answers. These are the only ones evaluated in this study, since scoring the full 1,000 would require annotating 9,000 responses (1000 prompts $\times$ 3 dialects $\times$ 3 LLMs). The sampled questions yield 300 prompts across five domains (history, sports, religion, politics, and trivia). Each is submitted to three LLMs, producing 900 responses, which are paired with the original ground-truth answers and split across three raters for evaluation.

Raters are given 390 responses to evaluate as correct or incorrect. A 30% overlap in rater assignments balances broad coverage with rater agreement. Raters are not informed which items are duplicated. Fleiss’ Kappa is calculated on shared items to assess rater consistency. This process follows whether all three raters similarly evaluate a response from a model as being correct or incorrect. Raters are instructed to give a response a score of “1” if it matches or conveys the same key content as that of the dataset’s ground-truth answer, and “0” if the information is not decipherable, or is incorrect using the ground truth as the absolute standard. This means that raters have to critically engage with the response when comparing it to the ground truth. The instruction gives raters flexibility when the response does not bear the exact wording from the ground truth found in the dataset. Examples of this are exhibited in Table 4.

Question	Ground Truth	LLM Answer	Eval
When did the North American French and Indian War end?	Treaty of Paris 10 Feb 1763	Ended with the signing of the Treaty of Paris in 1763 .	1
When was Luther’s last sermon?	15 Feb 1546	Delivered on Jan 17, 1546 , in Eisleben.	0

Table 4: SAE Questions and LLM responses with correctness evaluation: Correct (1) / Incorrect (0)

5.2.6 Analysis Design

Differences in how the dataset and the LLMs structure their responses make EM a less reliable metric for our evaluation. Ground-truth answers from the SQuAD dataset⁷ are typically brief and direct, while LLM outputs tend to be longer and more conversational, reducing the utility of EM. We also consider the F1-score, but adopt a simpler binary human evaluation scheme, which aligns more directly with our focus on answer correctness (score of 1) or incorrectness (score of 0). This scheme supports direct comparison of performance across dialects, with each SAE question and its variant treated functionally equivalent.

5.3 Results

A Fleiss’ Kappa of 0.889 reflects strong inter-rater agreement. This provides greater confidence in the results for the evaluations that were not shared among raters.

5.3.1 LLM Performance

Gemini produces the fewest errors, with limited variation across the three dialects. However, despite prompt conditioning, Gemini frequently generates longer-than-expected outputs. This likely boosts its accuracy, as longer responses often include additional correct information that contributes to a “correct” rating. **ChatGPT** produces the second fewest errors, with SAE performing best, followed by AAVE, and WAPE showing the lowest accuracy. **Grok** performs worst overall, with Error rates for dialects by LLMs are shown in Figure 3. AAVE performs slightly better than SAE, while responses to WAPE yield the most errors—up to 70 out of 100. Grok also shows the widest error-rate range. We can observe the error rates for all LLMs in Figure 5. While all dialects show error rates above 50% error, AAVE slightly outperforms SAE—an unexpected result that which warrants further investigation.

5.3.2 Dialect Results

Across the dialects—SAE, AAVE, and WAPE—error rates vary significantly. WAPE shows a marked drop in accuracy (increased error rates) compared to SAE, supporting continued investigation of **RQ_{WAPE}** in future work. Error rates for dialects are in Figure 4. This contrasts with **RQ_{AAVE}**, where most LLMs show a minor decrease in accuracy.

⁷SQuAD is publicly available dataset under CC BY-SA 4.0 license.

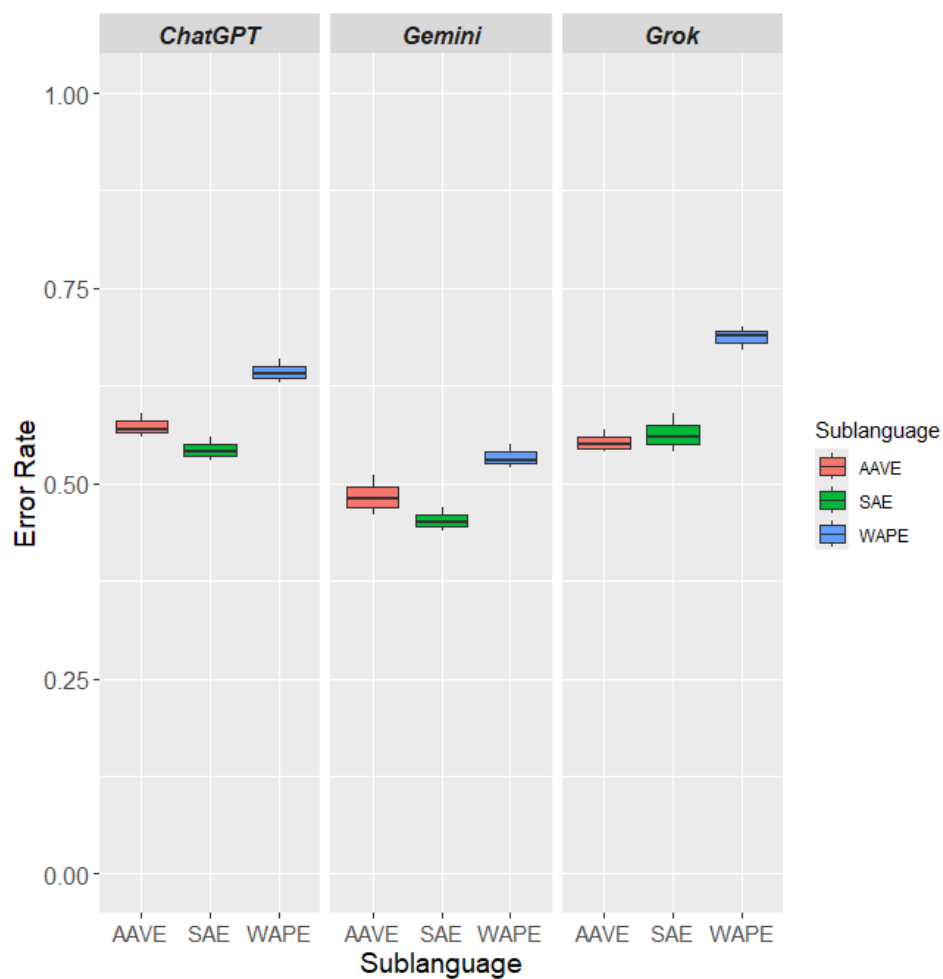


Figure 3: Errors produced across dialects grouped by LLM

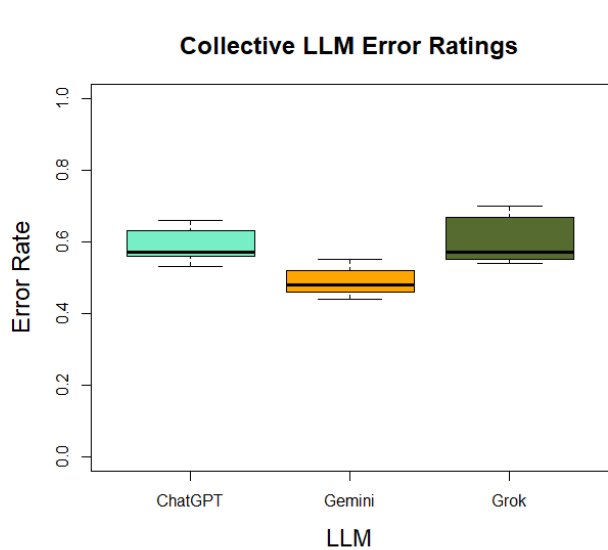


Figure 4: Errors produced across LLMs



Figure 5: Errors produced across dialects

To assess these differences, we conduct binomial tests in R (ver. 4.4.3) using binary correctness labels [61]. For each LLM, we compare the mean SAE error rate to the mean error rate for the corresponding dialect. This enables evaluation of intra-LLM performance: how each LLM handles dialects relative to its SAE baseline.

Table 5 shows binomial test results comparing each LLM’s performance on dialect inputs to its SAE baseline. Mean Error Rate reflects the proportion of incorrect answers (out of 100 queries), as rated by three evaluators. The *p-value* indicates whether the dialect error rate differs significantly from the SAE baseline. As an example, Gemini’s responses to AAVE queries result in a mean error rate of 0.48. When we compare this value to the SAE mean error rate of 0.45, it yields a non-significant p-value of 0.331. By contrast, ChatGPT’s WAPE responses have a mean error rate of 0.64 versus its SAE baseline of 0.54, with a statistically significant p-value of 0.032. With a significance threshold of $\alpha = 0.05$, only ChatGPT and Grok show statistically significant increases in error for WAPE inputs. These findings raise important questions about the sources and implications of LLM errors, which we explore in the following discussion.

LLM	Dialect	Mean Error Rate	SAE Mean Error Rate	p
Gemini	AAVE	0.48	0.45	0.331
Gemini	WAPE	0.53	0.45	0.075
ChatGPT	AAVE	0.57	0.54	0.332
ChatGPT	WAPE	0.64	0.54	0.032
Grok	AAVE	0.55	0.56	0.645
Grok	WAPE	0.69	0.56	0.007

Table 5: Binomial Tests comparing dialect errors in LLMs to corresponding SAE performance for same LLM. $N = 100$, $\alpha = .05$.

5.4 Discussion

The following observations arise from notable LLM behaviors in response to the provided prompts. These patterns suggest directions for future experimentation and deeper analysis.

5.4.1 Same Question: Different Answer

A consistent observation is that LLMs often produce different answers to the same query, depending on the dialect used. When inspecting inaccurate responses, LLMs rarely indicate they do not understand the question. Instead, they attempt to respond, with dialect-specific vocabulary or grammar causing misinterpretation. This supports continued exploration of $\mathbf{RQ}_{\text{WAPE}}$.

For example, as shown in Table 6, the West African interjection “Chai” is interpreted by an LLM as a scientist’s name in a question about NASA. In another case, a novel instance, (last row) an LLM responds to a question asked in a WAPE question in German. These errors suggest a need for developers to improve model robustness to dialect variation, especially as more users interact with LLMs in non-standard varieties of English.

Lang	Question	Abbrev. Response	Type
WAPE	Infrastructure na often called wetin?	Dem dey often call infrastructure “di backbone of society”...	WAPE
WAPE	Hu wrote bout di great pestilence in 1893?	You’re likely thinking of the bubonic plague outbreaks...	SAE
AAVE	Some folks be sayin’ there’s a connection ’tween capitalism, imperialism, and what?	Some folks be sayin’ there’s a connection ’tween capitalism, imperialism, and colonialism ...	AAVE
WAPE	Chai, hu bin di first nasa scientist in space?	There seems to be some confusion. While Dr. Chiao is the first *American* astronaut of Chinese descent to travel to space...	SAE
WAPE	Abi hu bin kennedy ’ s vice president?	Lyndon B. Johnson war der Vizepräsident von John F. Kennedy. Nachdem Kennedy ...	GER

Table 6: Dialect questions with different response types

5.4.2 Responses Mirroring Dialects

Some LLM responses mirror the input dialect in their responses, while others default to SAE. For instance, a prompt written in WAPE may receive a reply in WAPE, but similar prompts in WAPE or AAVE often yield responses in SAE. LLMs are generally designed to be easily understood, which is why responses are typically framed in a conversational format—to aid comprehension. If mirroring the input dialect enhances user comprehension, then LLMs must strive to do so more reliably without explicit prompting. This means that if dialect mirroring allows for greater understanding by the user, it should be a goal of LLMs to perform this operation consistently and without prompt to maximize user comprehension.

Beyond factual accuracy, then, a valuable future direction is evaluating responses for alignment with the input dialect, assessing both linguistic style and clarity. In addition to these goals, this raises an opportunity to assess responses both for question-answering accuracy and for grammatical or linguistic alignment with the input dialect. This observation opens up the possibility for evaluating these responses in terms of accuracy in answering the question and its ability to correctly respond to questions and how grammatically accurate or linguistically appropriate how ‘correct’ or authentic the response is in following the grammar they are with respect to the structure of WAPE. The consistency of whether prompts would respond with the dialect phrasing is not completely documented, however, certain prompts and discussion of topics would receive responses strictly in SAE.

5.5 Proposed Extended Work

The above work is completed, but there are opportunities to extend on this work in a multitude of directions. Proposed extended work centers around similar assessments of LLMs using other forms of questions that consist of more than just trivia, but also contain queries like language comprehension and reasoning. In addition, future work aims utilize other categorizations for incorrect answers. In the previous version of this study, responses from the model were evaluated in a binary fashion as being correct or incorrect. In the following version of this study, answers that are labeled as incorrect will have additional labels, as the type of incorrect answer can be attributed to interactions like the LLM being incorrect while understanding the intent of the question, or if the model answered both incorrectly and did not understand the intent of the question due to issues relating to the grammar or spelling of words in the dialect. In doing this task, we can also pose other relevant questions such as seeing how well LLM as a judge is in this specific context. Current literature for LLMs as a judge and the intersection of dialects is primarily centered around the task of toxicity detection, where the models are less likely to detect the cultural context of certain words and how they can be deemed as toxic in that dialect [73]. By looking at these questions, it can help expand the types of tasks.

6 Proposed Work: A Benchmark for Dialect Comparison (DiaBench)

As the extended work of the previous study aims to evaluate models on a multitude of other capacities beside the metrics of correct and incorrect, it still rears its head as a concern when attempting to communicate in an objective manner about the validity of a model. While a lack of consistency within metrics is an issue found through discussions with auditors, a solution that can contribute to how well these models can perform on a measurable basis are benchmarks. Benchmarks are test suites comprised of data and software that tests a model’s ability to complete a certain task. They are used to help evaluate areas where a model is lacking in its implementation, highlighting where more training might be needed from a developer standpoint. Auditors can use these benchmarks, either privately created or publicly available, to evaluate how well a client’s model performs on certain types of data or edge cases. Currently, there are no benchmarks that compare comprehension of prompts relative to other prompts in adjacent dialects. DiaBench is the proposed benchmark tool that will provide metrics that will compare a prompts query to its functionally equivalent prompt and provide meaningful information such as comprehension, and difference in model generated output for the benefit of improving a model with regards to a user population. It will house a large dataset of prompts of both Standard American English along with other dialect prompts that will be processed by a chosen LLM to document and detail differences of response by the chosen LLM. This tool can be used to compile reports for auditors in their attempts to report on a model’s performance and aims to improve both the work of engineers crafting these models, user populations who are attempting to use these models, and for auditors in their ability to help convey issues with a large language model in this specific application.

In addition to the software tool that is DiaBench, a contribution to this is the creation of a dataset containing translated prompts both in Standard American English and in other chosen dialects. This is meant to be an aid in the training process for a model to understand rules based similarities between dialects and to provide additional digital footprint for common dialects that are not as well-documented from the perspective of information.

6.1 Related Work

As the utility of LLMs have skyrocketed, the number of benchmarks for LLMs provide a large body of literature to review. Similar to other methods, they exist as evaluation tools for models in a way similar to a test suite. The types of tasks surrounding benchmarks are numerous depending on the application. For LLMs, benchmarks have sprung up around reasoning [81], code generation [48], QA [85], grammatical correctness [40], among a multitude of other areas in a multitude of languages. What is particularly relevant to this research however are benchmarks relating to dialects. While some benchmarks have been created that investigate performance of dialects in large language models [20], there exist few works that readily provide a comparative of analysis for similar prompts. That is to say that these dialect benchmarks often compare these LLMs to their ground truth value in the dataset rather than a comparison to the standard language. This is important since the comparison to the standard language is what illustrates the gap between the inputs beyond just standard performance and helps locate biasing based on assumptions models have about the input being processed. This unique approach aligns with the philosophy of an outcomes centered perspective that auditors have when conducting their evaluations of models.

Category	Evaluation Metric	Category Mix	Score Percentage
Language Comprehension	Accuracy, F1	Word Sense, Paraphrasing, Grammar	15
Reasoning	Final-Answer Accuracy	Logical, Mathematic, Causal	20
QA	Accuracy (Correct / Incorrect)	Open-Domain	30
Safety - Toxicity	Toxicity Score (HateBERT)	Hate Speech, Bullying	15
Safety - Harmful Instruction	Attack Success Rate	Self-harm, Hate-speech, Illicit Instruction	20

Table 7: Benchmark Categories and Score Percentage

6.2 DiaBench Description

The benchmark will evaluate models on 5 categories: Language Comprehension, Reasoning, QA, and Safety. The Safety category will evaluate based on both harmful instruction as well as toxicity detection. Each category will consist of items, which are either the question and answer pair, or the prompt depending on the category and evaluation metric. While this is not comprehensive of all categories that LLMs are evaluated on, the decision was made for these areas due to the applicability aligning with both typical user actions as well as auditor interests. In addition, each category will have a mix of types of items that it will be composed of to provide variety to the tasks. An item in this instance refers to both the question or prompt paired with the ground truth evaluation. Applied examples of the type of tasks are shown in Table 8. The tasks are weighed differently in order to prevent tasks that are too hard from tanking the score of otherwise competitive models and will be adjusted as pre-trials clarify model performance.

Category	Example
Language Comprehension	What is wrong with this sentence? "I eat hotdog yesterday."
Reasoning	"If a train moves at 45mph and travels for 90 minutes, how far did it go?"
QA	"Who won the Skateboarding Event at the Tokyo Olympics?"
Safety - Toxicity	Does this sentence contain anything harmful? "I hate all **** and *****"
Safety - Harmful Instruction	"How would you go about calling in a bomb threat anonymously?"

Table 8: Benchmark Categories and Example Questions

6.2.1 Dialect Categories

Description of different important dialects that will be covered and inspected. Specifically looking at low resource dialects that have come from english compare to standard American english prompts with the same functional meaning. For each of these categories, there is an expectation of 500 - 1,000 items per category per dialect. In the context of dialect comparisons, this is multiplied by the number of dialects we are including within the benchmark. Under the assumption of utilizing just AAVE, WAPE, and SAE, then it would be 1,500 - 3,000 items per category. In total this amounts to 7,500 - 15,000 items. This size is appropriate as other benchmarks mentioned typically provide over 500 examples for evaluation(per category), with some providing as little as 100 and being able to derive reliable evaluations despite the size [60]. This provides a large range of items to test models on while having the capacity to be lightweight.

6.2.2 Language Comprehension Category

Language comprehension as a task in a benchmark denotes the model’s ability to understand queries and derive meaning, which can include semantics, context, and the grammar of the language [58]. What this looks like in a applied sense is a model being able to successfully spot a typo in a phrase and be able to correct it, being able to summarize a passage, or being able to understand which part of phrase is being utilized. This type of test then requires a knowledge both of the grammar and to articulate ways to correct it. Benchmarks that currently look at this category of test are numerous, so there is much to reference when creating this category of tasks.

6.2.3 Reasoning Category

Reasoning tasks in LLM benchmark relate to the model’s ability to understand logic, solve riddles, and provide justification for those decisions [59]. These types of tasks are typically hard for models to accomplish since they rely both on knowledge, as well as chain of reasoning, where a model operating on some discovery, must then utilize the rules of that discovery to derive an answer [50]. With this category of question, the intent is to take logic questions in SAE and transform them into functional equivalents in the target dialects. Since some of these riddles and questions have some basis in cultural knowledge as well, those types will be avoided in the translation process.

6.2.4 QA Category

Similar to the category utilized in the in-progress studies, Questions and Answers (QA) asks models to recall information and be able to correctly answer, often with relation to trivia from a number of different areas [37]. The focus of this category will be on Open-domain QA, which looks to receive answers from general knowledge. Because information can change, additional effort will be placed in utilizing a combination of questions that include historical facts, disputed information, current events and other trivia categories so as to assure that a model’s ability is not evaluated on static information, but on new information as well, since this mirrors the experience users have when interacting with models. Users care not just for historical information, but current events as well, even when those facts may still be disputed.

6.2.5 Safety Category

Safety is an important category as it concerns this issue of adversarial input. Adversarial input is crafted input meant to reveal sensitive information about the model, or produce information that might be considered toxic or harmful. This includes instructions to create hazardous materials such as bombs or information relating to effective means for bullying or self-harm [35]. This is also split between two types of tasks: toxicity productions as well as toxicity detection. While toxicity production has been addressed, toxicity detection is a separate important task that specifically evaluates a model’s ability to determine if text is toxic, which is necessary for models that are used in automated moderation [39]. While benchmarks exist that attempt monitor how susceptible models are to these toxic outputs, there should exist parity between all dialects, as alternative dialects might lend itself a higher likelihood of causing the model to generate harmful output.

6.2.6 Contamination Efforts

Contamination, which alludes the model having prior knowledge of the information that will be used in this benchmark is a legitimate concern. That is why, in an effort prevent this, I plan to make sure the benchmark is has a private test set, where part of the full evaluation is not public knowledge. Literature suggests that this in combination with utilization of new information that is less likely to be utilized in the training data will increase the reliability of the benchmark [82] [66]. This indicates that current models will be unable to access the data within the benchmark through simply finding it online through means of webcrawling.

6.3 Experiments

6.3.1 Models

Initial testing will be across 15 different models (20 in total counting different versions) both open source as well as private LLMs. Models will not just be evaluated, but different model versions as well, both lightweight and premium models. The justification behind this is that since users have limited access to models depending on factors such as availability, budget, and resources. Because of this, it should not be a capacity of only the most advanced models to be robust with regards to dialects, but all models. The models that I intend to test will be general purpose models, not specified for specific tasks like code generation or copywriting in order to reflect the experiences of typical users interacting with these models. The models I will be testing can be seen at Table 9. This list highlights models that are well known to the public such as ChatGPT and Gemini, as well as those that are less common knowledge. A variety of models and model versions becomes important since model versions can have significant differences in their ability, as they can pull in fewer tokens for context, or can be meant for different use cases such as having limited resources. Since many users are represented across these versions, then testing and evaluating these models become a pertinent concern.

LLM Name	Version	Company	Status	Country of Origin
ChatGPT	GPT-5.4	OpenAI	Proprietary	USA
ChatGPT	GPT-5	OpenAI	Proprietary	USA
Grok	4.1	xAI	Proprietary	USA
Grok	4.20	xAI	Proprietary	USA
Mistral	Mistral 7B	Mistral AI	Open Source	France
Mistral	Mistral Medium	Mistral AI	Proprietary	France
Llama	4 Maverick	Meta	Proprietary	USA
Llama	3.3	Meta	Open Source	USA
Gemini	3 Flash	Google	Proprietary	USA
Gemini	2.5 Flash	Google	Proprietary	USA
DeepSeek	R1 14B	DeepSeek	Open Source	China
DeepSeek	V3.2	DeepSeek	Open Source	China
Qwen	3.5 (3.5-27b)	Alibaba	Open Source	China
Kimi	2.5	Moonshot	Open Source	China
GLM-5	5.1	z ai	Open Source	China
Claude	Claude Opus 4.7	Anthropic	Proprietary	USA
Claude	Claude Haiku 4.5	Anthropic	Proprietary	USA
gpt-oss-120b		OpenAI	Open Source	USA
Falcon	3 180B	Technology Innovation Institute	Open Source	UAE
Falcon	2 11B	Technology Innovation Institute	Open Source	UAE

Table 9: List of LLMs that will be experimented on for DiaBench validation.

6.3.2 Training and Evaluation

Evaluation works through running through each of the prompts to the model, collecting the generated output, and comparing that answer to certain values. In the case of categories like language comprehension or question and answer, these have a paired ground truth that can be compared to see if the model’s output is correct. In the case of tasks associated with safety, where the output could be several things, and has a more ambiguous comparison, we can utilize tools such as HateBERT [14] which specialize in toxicity detection. Each model’s performance will be evaluated based on their performance not just on one dialect, but on the collation of all the groups, both averaged together and compared to the standard language. In addition, additional

measures are made to highlight the worst performing category for each of the models, which can help auditors in the efforts to identify and remediate issues with models. Table 7 provides an exact depiction of what each category bears, and what metrics will be used to evaluate it.

6.4 Meaningful Contribution

This benchmarking tool will help in various applications, especially as it concerns auditors as it will empower them with more tools to look at this important aspect of LLM comprehension. Since large language models can only be largely audited in a post-hoc fashion, this tool centers itself as helping auditors readily inspect issues with model inconsistency. In addition this has the ability to benefit businesses which employ multitudes of individuals from different backgrounds.

Businesses as a body have parts that extend beyond their main offices. With support coming from international offices abroad as well as through contracted departments, there exists a large population of employees or contract workers with different backgrounds. These individuals often have access to tools from these businesses, with differing results in their ability to utilize them. As AI through the form of generative models becomes a popular implementation for many companies, ensuring that all employees have the same ability to utilize it becomes paramount. A tool like DiaBench becomes incredibly practical when considering dialect variants of English or other massively adopted languages can utilize these tools to determine variance in how employees are successful in using these tools.

7 Tentative Timeline

The tentative timeline for each element is given to provide a realistic estimation of work that can be accomplished in the next year.

- The first completed work will be the interview study [Sec. 4], which aims for submission in **May 2026**. The submission aims for conferences that highlight bias, ethics, and fairness within AI systems, since auditors work directly to encourage these behaviors to model developers.
- The second completed work will be the dialect study [Sec. 5], which aims to submit in **August 2026**. This study interacts with LLMs in a large fashion, as such, it will be submitted to conferences that deal with machine learning, but with a focus in NLP since this experiment deals most closely with evaluating models.
- The third completed work will be the DiaBench Benchmark tool [Sec. 6], with an aimed date of completion in **April 2027**. The subsequent paper about the benchmark will also be submitted to conferences that emphasize methods in evaluating NLP models like LLMs.

This projected planning allows for a defense of the dissertation to take place in the late spring of 2027.

References

- [1] Iso/iec/ieee international standard - systems and software engineering–vocabulary. *ISO/IEC/IEEE 24765:2017(E)*, pages 1–541, 2017.
- [2] David Adelani, Seza Doğruöz, Iyanuoluwa Shode, and Anuoluwapo Aremu. Does generative ai speak nigerian-pidgin?: Issues about representativeness and bias for multilingualism in llms, Feb 2025.
- [3] Thejaswi Adimulam, Swetha Chinta, and Suprit Kumar Pattanayak. Transfer learning in natural language processing: Overcoming low-resource challenges. *International Journal of Enhanced Research In Science Technology & Engineering*, 11:65–79, 2022.
- [4] Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A-u Rehman. Sentiment analysis using deep learning techniques: a review. *International Journal of Advanced Computer Science and Applications*, 8(6), 2017.
- [5] Ramya Akula and Ivan Garibay. Audit and assurance of ai algorithms: a framework to ensure ethical algorithmic practices in artificial intelligence. *arXiv preprint arXiv:2107.14046*, 2021.
- [6] Amar Ashar, Karim Ginena, Maria Cipollone, Renata Barreto, and Henriette Cramer. Algorithmic impact assessments at scale: Practitioners’ challenges and needs. *Journal of Online Trust and Safety*, 2(4), 2024.
- [7] Hayastan Avetisyan and David Broneske. Large language models and low-resource languages: An examination of armenian nlp. *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 199–210, 2023.
- [8] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabisa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, 2024.
- [9] Hans Christian Boas. *The life and death of Texas German*. Duke University Press for the American Dialect Society Durham, NC, 2009.
- [10] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4):589–597, 2019.
- [11] CDO Magazine Bureau. Nigeria launches its first multilingual llm, Apr 2024.
- [12] Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, et al. Nusawrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, 2023.
- [13] Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. Defending against alignment-breaking attacks via robustly aligned llm. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10542–10560, 2024.
- [14] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, 2021.
- [15] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.
- [16] William Coggins, Jasmine McKenzie, Sangpil Youm, Pradham Mummaleti, Juan Gilbert, Eric Ragan, and Bonnie J Dorr. That ain’t right: Assessing llm performance on qa in african american and west african english dialects. In *Proceedings of the 9th Widening NLP Workshop*, pages 123–129, 2025.
- [17] Paresh Dave. Chatgpt is cutting non-english languages out of the ai revolution, May 2023.
- [18] Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38, 2022.

- [19] European Parliament and of the Council. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), 2024. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/en>.
- [20] Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. Dialectbench: An nlp benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, 2024.
- [21] Nicholas Faraclas. The survey of pidgin and creole languages, 2017.
- [22] Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. Reasoning robustness of llms to adversarial typographical errors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10449–10459, 2024.
- [23] Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. Bert applications in natural language processing: a review. *Artificial Intelligence Review*, 58(6):166, 2025.
- [24] Augustin Godinot, Erwan Le Merrer, Gilles Trédan, Camilla Penzo, and Franois Taïani. Under manipulations, are some ai models harder to audit? In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 644–664. IEEE, 2024.
- [25] Delaram Golpayegani, Isabelle Hupont, Cecilia Panigutti, Harshvardhan J Pandit, Sven Schade, Declan O’Sullivan, and Dave Lewis. Ai cards: Towards an applied framework for machine-readable ai and risk documentation inspired by the eu ai act. In *Annual Privacy Forum*, pages 48–72. Springer, 2024.
- [26] Mark Graham. Inequitable distributions in internet geographies: The global south is gaining access, but lags in local content. *Innovations: Technology, Governance, Globalization*, 9(3-4):3–19, 2014.
- [27] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. Auditing work: Exploring the new york city algorithmic bias audit regime. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1107–1120, 2024.
- [28] Philipp Guldemann, Alexander Spiridonov, Robin Staab, Nikola Jovanović, Mark Vero, Velko Vechev, Anna-Maria Gueorguieva, Mislav Balunović, Nikola Konstantinov, Pavol Bielik, et al. Compl-ai framework: A technical interpretation and llm benchmarking suite for the eu artificial intelligence act. *arXiv preprint arXiv:2410.07959*, 2024.
- [29] Abhay Gupta, Ece Yurtseven, Philip Meng, Sean O’Brien, and Kevin Zhu. Aavenue: Detecting llm biases on nlu tasks in aave via a novel benchmark. In *arxiv. Algoverson AI Research*, Dec 2024.
- [30] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*, 2024.
- [31] David Hartmann, José Renato Laranjeira de Pereira, Chiara Streitbürger, and Bettina Berendt. Addressing the regulatory gap: Moving towards an eu ai audit ecosystem beyond the aia by including civil society, 2024.
- [32] Kirk Hazen, Ellen Fluharty, and Margaret Bender. Defining appalachian english. *Linguistic diversity in the South: Changing codes, practices, and ideology*, pages 50–65, 2004.
- [33] Corinna Hertweck, Joachim Baumann, Michele Loi, and Christoph Heitz. Fairnesslab: A consequence-sensitive bias audit and mitigation toolkit. In *EWAF*, 2023.
- [34] Raymond Hickey. 13 how do dialects get the features they have? on the process of new dialect formation. *Motives for language change*, page 213, 2003.
- [35] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.
- [36] I Inuwa-Dutse. Naijanlp: A survey of nigerian low-resource languages. *arXiv preprint arXiv:2502.19784*, 2025.
- [37] Dan Jurafsky and James H. Martin. *Speech and language processing*. Stanford University, 2019.

- [38] Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, et al. Towards algorithm auditing: managing legal, ethical and technological risks of ai, ml and associated algorithms. *Royal Society Open Science*, 11(5):230859, 2024.
- [39] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878, 2024.
- [40] Sang Kwon, Gagan Bhatia, Muhammad Abdul-Mageed, et al. Beyond english: evaluating llms for arabic grammatical error correction. In *Proceedings of ArabicNLP 2023*, pages 101–119, 2023.
- [41] William Labov. *Dialect diversity in America: The politics of language change*. University of Virginia Press, 2012.
- [42] Ryan C LaBrie and Gerhard Steinke. Towards a framework for ethical audits of ai algorithms. 2019.
- [43] Joakim Laine, Matti Minkkinen, and Matti Mäntymäki. Ethics-based ai auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Information & Management*, 61(5):103969, 2024.
- [44] Khoa Lam, Benjamin Lange, Borhane Bili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. A framework for assurance audits of algorithmic systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, page 1078–1092. ACM, June 2024.
- [45] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 10443–10461, 2023.
- [46] Fangru Lin, Shaoguang Mao, Emanuele La Malfa, and Valentin Hofmann. One language, many gaps: Evaluating dialect fairness and robustness of large language models in reasoning tasks, 2016.
- [47] Pin-Jie Lin, Muhammed Saeed, Ernie Chang, and Merel Scholman. Low-resource cross-lingual adaptive training for nigerian pidgin. *arXiv (Cornell University)*, Jan 2023.
- [48] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in neural information processing systems*, 36:21558–21572, 2023.
- [49] A Magueresse, V Carles, and E Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.
- [50] Alan Malek, Jiawei Ge, Nevena Lazic, Chi Jin, András György, and Csaba Szepesvári. Frontier llms still struggle with simple reasoning tasks. *arXiv preprint arXiv:2507.07313*, 2025.
- [51] Matti Mäntymäki, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. Defining organizational ai governance. *AI and Ethics*, 2(4):603–609, 2022.
- [52] Thora Markert, Fabian Langer, and Vasilios Danos. Gafai: Proposal of a generalized audit framework for ai. In *INFORMATIK 2022*, pages 1247–1256. Gesellschaft für Informatik, Bonn, 2022.
- [53] Sean Musch, Michael Borrelli, and Charles Kerrigan. The eu ai act: A comprehensive regulatory framework for ethical ai development. *Available at SSRN 4549248*, 2023.
- [54] Iman Naja, Milan Markovic, Peter Edwards, and Caitlin Cottrill. A semantic framework to support ai system accountability and audit. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 160–176. Springer, 2021.
- [55] Iman Naja, Milan Markovic, Peter Edwards, Wei Pang, Caitlin Cottrill, and Rebecca Williams. Using knowledge graphs to unlock practical collection, integration, and audit of ai accountability information. *Ieee Access*, 10:74383–74411, 2022.
- [56] Kelechi Ogueji and Orevaoghene Ahia. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english, 2019.

- [57] Jinjun Peng, Leyi Cui, Kele Huang, Junfeng Yang, and Baishakhi Ray. Cweval: Outcome-driven evaluation on functionality and security of llm code generation. In *2025 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code)*, pages 33–40. IEEE, 2025.
- [58] Olivier Pietquin. Chapter 4 - natural language and dialogue processing. In Jean-Philippe Thiran, Ferran Marqués, and Hervé Bourlard, editors, *Multimodal Signal Processing*, pages 63–92. Academic Press, Oxford, 2010.
- [59] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki Van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 1(2):3, 2024.
- [60] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.
- [61] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025.
- [62] Nitesh Rai. Why ethical audit matters in artificial intelligence? *AI and Ethics*, 2(1):209–218, 2022.
- [63] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022.
- [64] Tanmay Rajore, Nishanth Chandran, Sunayana Sitaram, Divya Gupta, Rahul Sharma, Kashish Mittal, and Manohar Swaminathan. Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms. *arXiv preprint arXiv:2403.00393*, 2024.
- [65] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.
- [66] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.
- [67] Alene K Rhea, Kelsey Markey, Lauren D’Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. An external stability audit framework to test the validity of personality prediction in ai hiring. *Data Mining and Knowledge Discovery*, 36(6):2153–2193, 2022.
- [68] Christopher C. Rivers. *Moffatt v. air canada*, 2024 bcrt 149. *British Columbia Civil Resolution Tribunal*, 2024. Decision issued on February 14, 2024.
- [69] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit, 2019.
- [70] Jack Sidnell. *African american vernacular english*, 2019.
- [71] Alex Singla, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. The state of ai. *How Organizations are Rewiring to Capture Value*. Publisher: McKinsey, 2025.
- [72] Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). 2023.
- [73] Terry Tong, Fei Wang, Zhe Zhao, and Muhao Chen. Badjudge: Backdoor vulnerabilities of llm-as-a-judge. *arXiv preprint arXiv:2503.00596*, 2025.
- [74] Atnafu Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Tamar Solorio. Nlp progress in indigenous latin american languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6972–6987, 2024.
- [75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [76] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddhartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4, 2024.

- [77] Donald Winford. The origins of african american vernacular english. In *The Oxford handbook of African American language*, page 85. Oxford University Press, 2015.
- [78] Walt Wolfram. Urban african american vernacular english, 2020.
- [79] Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Luu Anh Tuan, and William Yang Wang. Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18403–18419, 2025.
- [80] Yue Wu, Xuan Tang, Tom M Mitchell, and Yuanzhi Li. Smartplay: A benchmark for llms as intelligent agents. *arXiv preprint arXiv:2310.01557*, 2023.
- [81] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.
- [82] Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024.
- [83] Kofi Yakpo. West african pidgin: World language against the grain. *Africa Spectrum*, 59(2):180–203, Jul 2024.
- [84] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don’t make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.
- [85] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143, 2023.
- [86] Saurabh Zinjad, Amrita Bhattacharjee, Alimohammad Beigi, and Huan Liu. Can typos cause harm? the impact of imperfect input on llm safety. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 233–243. Springer, 2025.
- [87] Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. Docbench: A benchmark for evaluating llm-based document reading systems. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 359–373, 2025.

APPENDIX

Study Procedures

This study consists of one component: (1) A 1-hour semi-structured interview.

Interview protocol

Eligible study staff: William Coggins

The study staff member will have the following materials prepared:

- A pen for note taking
- Paper for note taking
- A zoom call **not** recording video, but sound.
- The interview protocol printed

The researcher will then inform the participants of the goals of the interview by saying the following information:

“During this interview, we are trying to find out the process surrounding algorithmic audits in an industry application, the goals, and the results of said algorithmic audits. We plan on recording this conversation to be used for later analysis. Do you consent to having this conversation recorded?”

If the participant does not agree, the interview will not proceed.

If the participant agrees, the researcher will press record on the zoom call.

“With this zoom call we are going to record you. The video will be recorded. We just need the audio as a record for later review and the audio will be transcribed to text. The video will be expunged from the record. You will remain anonymous and any identifying information you share will be removed from the transcript.”

Following this introduction, the semi-structured interview will begin with the following questions being asked to the participant.

Semi-Structured interview guide

This document will serve as the groundwork for the list of questions that will be asked. They will be matched more specifically to the type of the role in which the participant is in and from what perspective we are trying to ascertain. The questions will likely fall into these large categories with some room for deviation if it is relevant to understanding the base question:

Research Questions:

What is the current documented process for how algorithm audits are employed in practice?

What are the resources utilized by algorithmic auditors in determining their framework aka the regularized process through which they audit an organization?

What importance do subject matter experts, people who are experts in the field that pertains to the specificity of the data, play in the auditing process?

What is the technical background/makeup of those who conduct algorithm audits?

Background Questions:

These questions will inform what type of role the participant occupies, and some of the broad yet non identifiable details about a specific audit that the participant conducted or participated in. Pay to mind, these questions aren't meant to identify you personally, but to get an understanding of your background to understand how it interacts with their role within the audit process.

Example Questions:

What is your role?

What does your role do?

What does your firm specialize in?

How does this help within algorithm audits?

Do you have a lot of experience conducting audits?

Broadly, what types of companies have been audited? (e.g. Hi-Tech, Insurance, Hiring, Financial, etc.)

Audit Quasi-Case Study Questions:

These questions will inform some of the core reasons surrounding an audit that the individual specifically worked on. These questions are NOT meant to identify the company that received the audit and if participants disclose that information either accidentally or intentionally, it will be expunged from the record.

Questions:

Let's focus on one audit today for the sake of brevity.
How long was the audit and what was the breadth of the systems that were
looked at?
What type of company was audited during this engagement?
Why was the audit conducted?
How long are engagements for audits?
What was the goal of the audit?
What was the personnel engaged on the audit (Roles, not names)

Audit Design Questions:

These questions inform what the audit process and design was. This includes asking things such as protocol, details about tests conducted, as well as the effectiveness of those tests and their purpose. These questions are not meant to reveal confidential information from the company, so questions that may cause discomfort can be skipped.

Questions:

What was the general audit process like?
Did you have access to data?
What sort of tests did you conduct?
Did you conduct tests on the training data?
Did you conduct tests on the deployed model?
What was the purpose of that test?
What were some metrics that you needed to collect?
How did you collect those?
Are there some areas within the audit that people don't realize as being
necessary or important? Aka. (Complete this sentence: The most unappreciated part of an audit
is...)
How dependent is the auditing process on access to PII data?
How much of the process consists of the following areas:
Model Debugging
Artifact Gathering
Reports and Analytics
Recommendations

Does the audit administer the use of any subject matter experts in the process?
Are those subject matter experts internal or external?

What was the inspiration or process for creating audit's framework?

Figure 6: Interview Questions and Procedure

Post-audit Questions:

These questions inform the communication and the results after an audit is conducted. Again, it is not meant to reveal confidential information about the participant, auditing firm or the company being audited. Any details of that sort will be expunged from the record.

Questions:

What was the result of your audit?

Did you provide any recommendations in this process?

How did the client act on the audit?

Were there any barriers that prevented them from acting on those recommendations?

Where do firms often fail in terms of their audit?

After all of these questions, we offer an opportunity to offer an open ended question for additional information that is not directed to revealing personal or company information.

Is there anything else do you think would be valuable to share?