Some Dialects Are More Equal Than Others: Non-Prestigious Arabic Dialectal Bias in LLMs

Mai Mohamed Eida[1], Ryan Dolan[1], Paul de Nijs[1], Jonathan Dunn[1]
[1]University of Illinois Urbana Champaign
{maimm2, ryanpd3, pdenijs2, jedunn}@illinois.edu

Abstract

Previous work on Egyptian Arabic in NLP has focused largely on the prestigious Cairene Egyptian Arabic (CEA) dialect, resulting in a lack of representation for the less prestigious Sa'idi Egyptian Arabic (SEA) dialect both in LLM and resource development. Does this lack of representation influence an LLM's view of the acceptability of SEA (upstream), and does an upstream bias against SEA lead to worse performance (downstream)? We investigate the upstream effect of SEA dialectal features on LLM preferences in a Targeted Syntactic Evaluation (TSE) task which reveals a significant bias against SEA across multiple LLMs. We then analyze the effect of these same features on downstream model performance on MMLU benchmarks and show that models experience a degradation in performance when presented with SEA. This work highlights the need for further exploration on how sub-dialectal variation impacts language technologies.

1 Introduction

Arabic dialectal variation remains a continuous challenge in Arabic Natural Language Processing (NLP). With over twenty-six dialects, and hundreds of sub-dialects (Behnstedt and Woidich, 2013; Versteegh et al., 2006–2009), a central goal of the field has been to model dialectal varieties of Arabic across language technologies, including Large Language Models (LLMs). However, accounting for such diverse dialectal variation is difficult and does not necessarily account for sub-dialectal variation within each country, and therefore, in this process, much of sub-dialectal variation is lost in practice.

Egyptian Arabic (EA) provides a clear case for supporting marginalized sub-dialectal varieties in Arabic language technologies. Egyptian Arabic is the most thoroughly modeled dialect in Arabic NLP, but it has tacitly become synonymous with "Educated Cairene Egyptian Arabic", the most prestigious dialect (Bassiouney, 2018) spoken by middle class educated Egyptians in Cairo and surrounding urban cities (Abdel-Malek, 1972; Badawi, 1973). This variety represents only one of the five groups of Egyptian Arabic sub-dialects, which vary across phonology, morphology, syntax, semantics, and lexicon (Behnstedt and Woidich, 1985; Badawi, 1973) and also vary culturally from Cairo. In this paper, we focus on the second most spoken dialect in Egypt, Sa'idi Egyptian Arabic (SEA) which is not linguistically represented within the dialectal landscape in current Arabic language technologies (Mohamed Eida et al., 2024).

The focus on Cairene Egyptian Arabic (CEA) is consistent with accessibility and availability of resources for prestigious dialects. Dialect attitudes are mediated by socioeconomic status, educational access, and language ideology (Labov, 2006). Speakers of prestigious dialects have the means to participate in digital spaces and contribute linguistic data which increases their representation in language technologies. On the other hand, speakers of marginalized dialects may avoid using their dialects online, even when access is available. This produces a self-reinforcing cycle in which language technologies prefer certain dialects while excluding others due to lack of linguistic resources.

Previous work has demonstrated that SEA is not well represented in online Egyptian Arabic corpora (Mohamed Eida et al., 2024). However, whether this lack of representation actually results in upstream or downstream bias against SEA in model behavior has not yet been demonstrated. It is possible that, despite the scarcity of SEA representation in online corpora on which LLMs are trained, the internal representations learned by LLMs are generalized enough that they are insensitive to dialect differences. If so, models may exhibit comparable downstream performance on SEA and CEA, suggesting that the scarcity of SEA resources does not necessarily translate into a poorer experience

Proceedings of The Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026), pages 48–64
October 28-29, 2026 ©2026 Association for Computational Linguistics

for SEA speakers.

This paper addresses two research questions. **First, do LLMs replicate the same bias humans have towards prestigious and non-prestigious sub-dialects in their upstream representations?** If so, this supports previous work on the lack of SEA representation in pre-training data in Arabic NLP and that such lack of representation can result in an upstream bias. In experiment 1, our results confirm there is a bias in monolingual, bilingual, and multilingual LLMs against SEA, with internal representations favoring CEA or Modern Standard Arabic (MSA) over SEA in almost every single case. **Our second research question addresses the implications of upstream bias: does this upstream bias impact the performance of LLMs on downstream tasks?** Specifically, we examine performance on Arabic Massive Multitask Language Understanding (MMLU) benchmarks. If there is no impact on downstream tasks, this may shift the focus of future work to other necessary tasks such as cultural awareness of sub-dialectal variation rather than curating SEA-specific corpora. The results from our second experiment show that downstream bias varies more across models, but still shows consistent favor towards CEA overall[1].

2 Background

2.1 Language Attitudes

Language attitudes are structured by social prestige and power (Labov, 2006). Dialect choice is not neutral. It is shaped by socially embedded evaluations that influence both perception and production. Speakers converge toward or diverge from others' speech depending on social attitudes and group identity (Giles, 1970). Perceptual dialectology shows that listeners systematically associate dialects with social attributes, assigning traits such as intelligence, competence, and professionalism to prestigious varieties, while attributing negative characteristics such as lack of education or credibility to marginalized dialects (Giles, 1970; Lambert et al., 1960; Preston, 1993).

This extends to Arabic dialects, with language attitudes spanning sub-dialects of many Arabic dialects (Aldosaree, 2016; Eltouhamy, 2015; Hachimi, 2015). As the most prestigious dialect, Cairene Egyptian Arabic (CEA) is socially constructed as educated, modern, and authoritative, while Sa'idi Egyptian Arabic is stereotyped as rural, uneducated, and socially marginal (Bassiouney, 2017; Haeri, 1997). This bias affects how speakers are evaluated and treated in education, employment, and institutional settings, with speakers of prestigious dialects being perceived as more competent and credible than speakers of stigmatized varieties. However, do LLMs exhibit a similar bias toward prestigious dialects mirroring existing human bias?

2.2 Dialectal Bias in LLMs

Recent research has shown that LLMs frequently demonstrate bias against certain language varieties across many different languages. Lin et al. (2025) show that LLMs systematically display a high degree of unfairness (performance gaps) and brittleness (sensitivity to input perturbations) when responding to reasoning queries in African American English (AAE). Similar work by Pan et al. (2025) shows that such bias also exists for multiple-choice reasoning benchmarks and extends to other English varieties such as Chicano, Appalachian, Southern, Indian, and Singaporean English. These works crucially demonstrate that even in tasks such as code generation or queries for factual world knowledge which should not reasonably vary across dialects, LLMs show a systematic bias.

Outside of benchmark performance bias, Fleisig et al. (2024) use qualitative evaluations from native speakers of 10 different varieties of English to show that GPT-3.5 and GPT-4 often use demeaning and stereotypical language when responding to varieties other than Standard American and Standard British English. They show that the models often misuse or exaggerate dialectal features in a way that is unnatural and perceived as mockery. Hassan et al. (2025) additionally show that LLMs demonstrate a systematic bias against AAE when used for sentiment analysis tasks, with AAE language being much more likely to be identified as negative. Outside of English, Bui et al. (2025) show that systematic bias also exists for non-standard dialects of German such as Bavarian and Alemannic varieties, with non-standard varieties more likely to be associated with negative adjectives in an adjective-association task. There is also work on Bengali and dialectal bias in this low-resource language based on the minorities/power structures (Sadhu et al., 2025), showing LLM production exhibit significant gender and religious bias in Bangala. In all of this research, LLMs do exhibit bias against the less prestigious varieties, and this impacts downstream

[1]All data, code, and prompts are available on https://github.com/maimm2/EA_Dialectal_Bias.

performance. Is this also the case for Egyptian Arabic?

3 Experiment 1: Do LLMs have a preference for one dialect?

Since SEA is highly under-represented compared to CEA in existing Egyptian Arabic datasets and tools (Mohamed Eida et al., 2024), is there a preference for one dialect over the other due to the scarcity of data? If so, then we expect LLMs to prefer CEA to SEA by finding CEA sentences more probable than an equivalent SEA sentence in a minimal pair.

3.1 Methodology

Models We select five pretrained models to test upstream preferences: one monolingual Arabic model specifically trained on both MSA and Tweets in Dialectal Arabic (DA) (henceforth "monolingual", Jasmine-350M: Billah Nagoudi et al. 2023), two English-Arabic bilingual models trained on a variety of Arabic dialects (henceforth "bilingual", Jais-2-8B & Jais-1-13B: Anwar et al. 2025; Sengupta et al. 2023), one multilingual model trained on generic Arabic data (henceforth "multilingual", Llama 3.1-8B: AI@Meta 2024), and one multilingual model continuously-pretrained specifically on Egyptian Arabic (henceforth "EA", NileChat-4B: Shang et al. 2025)[2].

Task To test upstream preferences, we employ Targeted Syntactic Evaluation (TSE, Linzen et al., 2016). In this approach, the probabilities of two sentences in a minimal pair are compared to decide which of the minimal pair of dialects the model prefers. We create minimal pairs between CEA and SEA by sampling items from an SEA corpus (Mohamed Eida and Habash, 2025) and changing features specific to SEA to their CEA equivalents. These features include phonetic, morphological, orthographic, and lexical variations. For most minimal pairs, only one feature at a time is changed, although we also include a "mixed" condition in which multiple features are changed at once.

Minimal Pair Construction We additionally create an ungrammatical baseline (UG) in where we replace the valid Arabic features such as morphemes and lexical items with meaningless nonce forms

[2]NileChat-4B is only continuously-pretrained on Egyptian Arabic text, but is later instruct-tuned with both Egyptian Arabic and English.

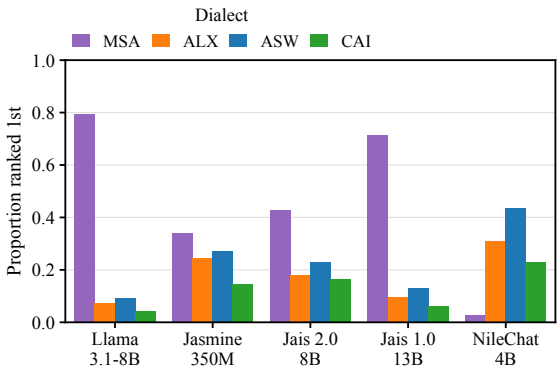

Figure 1: Proportion of dialect preference by model in MADAR, defined as being the most preferred variation in the TSE task. MSA is consistently most preferred, followed ASW (Aswan), ALX (Alexandria), then CAI (Cairo). With the exception of NileChat-4B, MSA is most preferred. NileChat-4B was precontinuously trained on Egyptian Arabic and therefore, preferring EA to MSA is expected.

that are less likely to correspond to an Arabic variety. For example, the CEA future marker هـ (h) and its SEA counterpart ح (ʕ) were replaced with د (d), which does not function as a future tense marker in Arabic dialects. We consistently replaced د (d) as future tense in every applicable case thereafter. This ensures that the modified forms preserved superficial characteristics such as length while removing their grammatical interpretation, and we control for as much variable change as possible. The UG forms generated matched the morphological, phonological, and orthographic features in length and script, as illustrated in Appendix B in Table 4. Lexical features were altered with +/-4 letters depending on how easy it was to create a nonce word. This accounts for the possibility of an LLM considering one or two letter changes in variants a typological error ("typo"). The UG baseline allows us to identify whether the model prefers CEA over SEA because it finds SEA to be purely ungrammatical, or because it considers SEA as a valid but dispreferred dialect. This yields a total of 134 minimal pairs, with samples illustrated in Table 3 in Appendix B.

Evaluation To compute sentence probabilities, we follow Leong and Linzen (2023) in our implementation of TSE by adding the (negative) log probabilities of each token in a given sentence (conditioned on previous tokens) to obtain the model's "judgment" for that sentence. We then normalize

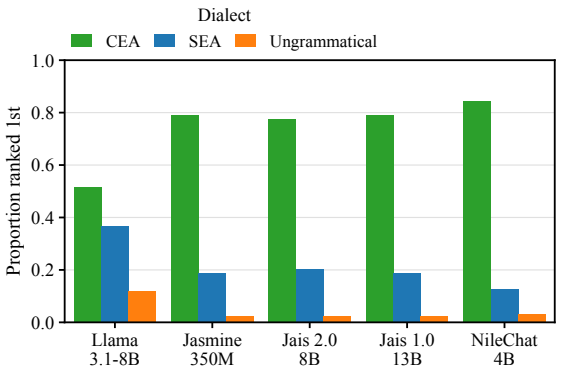

Figure 2: Proportion of dialect preference by model, defined as being the most preferred variation in the TSE task. CEA is consistently most preferred, followed by SEA and UG.

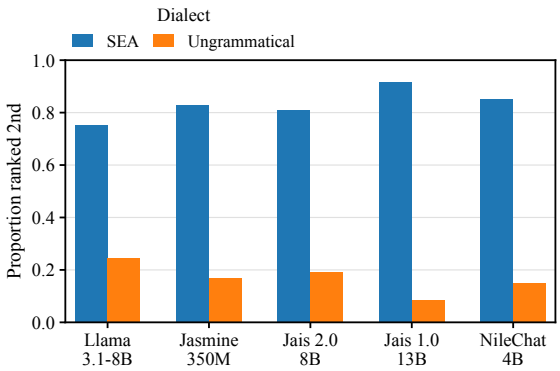

Figure 3: Second-choice model preferences when CEA is ranked most probable. All models strongly prefer SEA to the Ungrammatical baseline, indicating that SEA is viewed as legitimate, but dispreferred, variation.

this value by the length of the sequence to ensure that longer sentences are not artificially less probable.

As MSA remains the dominant Arabic variety in available NLP resources and is substantially better represented than dialectal Arabic in LLM training data (Mousi et al., 2025), we first check if MSA is the most preferred dialect over Egyptian Arabic (EA). To do so, we use the MADAR (Bouamor et al., 2018) parallel Arabic corpus and extract Arabic minimal pairs between MSA and the three Egyptian cities in the corpus: Cairo (CAI), Alexandria (ALX), and Aswan (ASW). We hypothesize that MSA will be the most preferred variety, given both its prestige and prevalence in available training data. Then, we examine which of our three CEA, SEA, and UG variants is most preferred by the model in each minimal pair. We then determine pairwise preferences between CEA and SEA and SEA and the ungrammatical baseline (UG). If the model exhibits a bias against SEA, we expect that the CEA variant will be consistently ranked the most probable sentence. If little or no bias exists, we expect that CEA will be ranked over SEA in roughly equal proportion to SEA being ranked over CEA. Furthermore, if CEA is consistently ranked higher than SEA, and SEA is consistently ranked higher than UG, this would indicate that the model recognizes SEA as grammatical, but still disprefers it to CEA. If there is no consistent preference for SEA over UG, this indicates that the model views SEA as fully ungrammatical.

3.2 Results

Is MSA more preferred than EA? As illustrated in Figure 1, MSA is the most preferred variety of Arabic when compared to Egyptian Arabic. Therefore, our hypothesis that MSA, as both the most prestigious and the most represented variety in training data, would be preferred holds across all models except NileChat-4B, where Egyptian Arabic emerges as the most preferred variety. Given that NileChat-4B is the only model continuously pre-trained specifically on Egyptian Arabic, its preference for Egyptian Arabic over MSA is expected.

Do LLMs prefer CEA over SEA? The ranking preferences of each model are given in Figure 2. In all models, a clear preference is shown for the prestigious CEA dialect over the less prestigious SEA dialect. The preference is very clear in the Arabic monolingual (Jasmine-350M), Arabic-English bilingual (Jais-1-13B and Jais-2-8B), and Egyptian Arabic (EA) model (NileChat-4B), with CEA being preferred around 80% of the time. This preference notably drops in the multilingual model (Llama-3.1-8B) to only 50%. A similar pattern is found for the ungrammatical baseline UG, which is found most preferable only around 2% of the time for all models except Llama-3.1-8B (11%). Therefore, LLMs prefer CEA over SEA and display an upstream bias against SEA.

Do LLMs consider SEA a legitimate variety? Having demonstrated that all models prefer CEA to SEA, we now ask whether this is because SEA is viewed as completely ungrammatical, or whether the model recognizes the SEA variants as being

legitimate, but still dispreferred. The results in Figure 2 suggest that SEA should be viewed as legitimate given that it is ranked above CEA in a non-negligible number of cases. Indeed, we find that in cases where CEA is ranked to be the most probable, SEA is preferred over the ungrammatical baseline in around 90% of cases (Figure 3). Thus, it is not the case that SEA is viewed as completely ungrammatical; rather, the models are simply biased towards CEA.

The absolute rankings of negative log probabilities clearly demonstrate a preference hierarchy of CEA > SEA > UG across all models, with only Llama-3.1-8B showing a somewhat weaker preference. Thus, in the majority of cases, CEA is indeed preferred over SEA, but SEA dialectal features are recognizably legitimate variations such that SEA is still systematically preferred over the UG ungrammatical baseline.

To what degree do LLMs prefer SEA over the ungrammatical baseline? Having established this preference hierarchy and the fact that SEA is not considered purely ungrammatical, we now turn to the question of whether SEA is considered closer to CEA overall than to the ungrammatical baseline. We do so by comparing the distribution of the *differences* in sequence probabilities between CEA and SEA ($P(Seq_{CEA}) - P(Seq_{SEA})$) and SEA and the ungrammatical baseline UG ($P(Seq_{SEA}) - P(Seq_{UG})$) for each minimal pair using a paired t-test. This allows us to determine the degree of preference; if, on average, $P(Seq_{CEA}) - P(Seq_{SEA}) < P(Seq_{SEA}) - P(Seq_{UG})$, this implies that the model considers SEA to be overall closer to CEA than the ungrammatical baseline. If the converse is true, this implies that the model considers SEA to be closer to the ungrammatical baseline than to CEA, even if it consistently ranks SEA higher.

The paired t-test is only significant for Jais-1-13B ($t = -2.59, p = 0.03^*$), which indicates that for all other models, SEA is not closer on average to CEA than to complete ungrammaticality, or vice versa. This can also be seen visually in appendix A. This shows that SEA interestingly occupies a sort of midpoint between CEA and UG in terms of grammaticality.

Does the number of characters changed matter? Finally, we assess whether the upstream bias exhibited by the models correlates with the scale of the grammatical feature change as measured by character edit distance. We correlate minimum edit distance and the difference in probabilities examined in the previous section. Edit distance was positively correlated and statistically significant with the probability difference CEA - SEA ($r = 0.18, p \approx 0.00$) and CEA - UG ($r = 0.17, p \approx 0.00$) for all models except for Llama-3.1-8B. However, SEA - UG did not show a significant correlation ($r = 0.009, p = 0.8$) with edit distance for any model. This shows that the magnitude of the change does affect the degree to which CEA is preferred over SEA, but that UG is heavily dispreferred regardless of the number of characters changed.

In summary, the consistent ranking of CEA over SEA shows a strong bias against SEA in upstream model preferences, answering our first research question. It is especially noteworthy that the minimal pairs were constructed using samples from an SEA corpus, changing only one overt SEA feature to its CEA counterpart. Despite this SEA source context, models still consistently prefer the CEA variant once that single feature is changed. In other words, the models prefer a CEA feature in an SEA context even over an SEA feature in an otherwise SEA context. Thus, the preference is strong enough that the model prefers an out-of-place prestigious dialectal feature to a well-suited non-prestigious dialectal feature. On the other hand, SEA is ranked higher than the ungrammatical baseline, demonstrating that SEA is a viable but dispreferred dialect occupying a midpoint between the prestigious CEA dialect and total ungrammaticality.

4 Experiment 2: Does Upstream Bias Impact Downstream Performance?

The results in Experiment 1 reveal a persistent upstream dialectal preference for CEA over SEA. However, does upstream bias actually impact downstream performance? Experiment 2 aims to answer this question. If so, this would indicate that marked sub-dialectal variation affects downstream performance, and emphasizes the need for sub-dialectal variation in resource development, model development, and language technologies overall.

4.1 Methodology

We adapt Pan et al. (2025)'s methodology, which evaluates the extent to which grammatical features associated with non-standard English varieties influence LLM performance on benchmarks such as Massive Multitask Language Understanding

(MMLU), Boolean Questions (BoolQ), and Science Questions (SciQ). We extend this methodology to three Arabic varieties: MSA, CEA, and SEA using the same set of features as in Experiment 1. We exclude lexical and orthographic variations because they are more difficult to consistently apply across all benchmarks. For each feature, we use a regular expression to first select all benchmark questions in which the feature occurs. We then create a variant of the question and multiple-choice answers with the target feature changed to its equivalent in MSA, CEA, and SEA. As in Experiment 1, we also produce an ungrammatical control variant (UG) as a baseline following the same rules of length, script and consistency. This transforms each original item into four versions corresponding to MSA (ArabicMMLU only), CEA, SEA, and UG conditions.

Benchmarks For this experiment, we focus on three MMLU benchmarks specifically developed for Arabic varieties. We select one MSA MMLU benchmark, ArabicMMLU (Koto et al., 2024), and two Dialectal Arabic (DA) benchmarks, AraDiCE-ArabicMMLU-EGY (Mousi et al., 2025) and DialectalArabicMMLU EGY (Altakrori et al., 2026). For the latter two benchmarks, we extract the Egyptian Arabic items only. This creates two conditions: one in which the original questions are fully in MSA, and one in which the original questions are in Egyptian Arabic. Having both types of conditions is important, because the model may behave very differently when faced with an SEA feature in MSA versus an SEA feature in Egyptian Arabic. Additionally, because MSA is a standard/neutral variety and the most preferred compared to EA as illustrated in Experiment 1, we can safely add CEA and SEA features in isolation from one another without the potential confound of an item already being written in CEA. By using MSA as the source condition in at least one benchmark, we can more directly isolate the effect of introducing CEA and SEA features while holding all other aspects of the item constant. The two Egyptian Arabic benchmarks are still included because both CEA and SEA are primarily spoken dialects, and thus provide a more natural context in which SEA features would actually occur. We list the benchmarks and the 24 extracted features in Table 4 in Appendix B.

Models We evaluate benchmark performance on three of the five models examined in Experiment 1: Llama 3.1-8B (AI@Meta, 2024), Jais-2-8B (Anwar et al., 2025), and NileChat-4B (Shang et al., 2025). Jasmine-350M (Billah Nagoudi et al., 2023) was excluded because it was not instruct-tuned, and Jais-1-13B (Sengupta et al., 2023) was excluded because the results of Experiment 1 were extremely similar to those of the newer Jais-2-8B.

Evaluation We initially evaluate each grammatical feature in isolation, modifying a single feature at a time and measuring benchmark accuracy across the MSA, CEA, SEA, and UG conditions. We then introduce a "mixed-features" condition, in which two or more grammatical features are changed simultaneously within the same benchmark item. Some feature perturbations involve only minimal changes, sometimes only a single character. It is possible that models treat these forms as spelling variations or noise rather than meaningful dialectal distinctions. This is corroborated by the positive correlation between edit distance and difference in upstream preference in Experiment 1 between CEA and SEA. Therefore, evaluating both isolated and combined feature substitutions we can determine whether minimal dialectal perturbations affect model accuracy to the same degree as larger dialectal perturbations.

5 Results

Single Feature Accuracy Overall The results for the single feature condition are given in Table 1. We first report the accuracy across all items on each benchmark for each model and dialect condition ("Overall Accuracy"). The results show that across almost all models and benchmarks, the model achieves higher accuracy when the questions and answers are in CEA than when they are in SEA. The magnitude of difference is not particularly consistent across models, with different models being impacted more across different benchmarks. Furthermore, the original dialect of the question does not appear to impact the degree of difference, with ArabicMMLU (MSA) having a similar margin of difference as compared to AraDiCE-ArabicMMLU-EGY and DialectalArabicMMLU (Egyptian Arabic).

Although higher accuracy is observed on CEA as compared to SEA, the margin is often small (and in fact there is one marginally positive case in favor of SEA: Jais-2-8B in the DialectalArabicMMLU benchmark). This is consistent with the results of Pan et al. (2025), who observe that non-prestigious English varieties degrade performance

		ARABIC MMLU MSA (2024) % (n=9689)			ARDICE MMLU EGY (2025) % (n=13400)			DIALECTALARABICMMLU (2026) % (n=6739)		
		JAIS 2.0 (8B)	LLAMA 3.0 (8B)	NileChat (4B)	JAIS 2.0 (8B)	LLAMA 3.0 (8B)	NileChat (4B)	JAIS 2.0 (8B)	LLAMA 3.0 (8B)	NileChat (4B)
Overall Accuracy	MSA	**65.86% (+0.46%)**	43.60% (-0.21%)	46.39% (+0.52%)	-	-	-	-	-	-
	CEA	65.39%	**43.80%**	**45.88%**	**57.45%**	**37.43%**	**43.54%**	42.25%	**34.59%**	**43.10%**
	SEA	64.28% (-1.11%)	43.03% (-0.77%)	45.29% (-0.59%)	56.05% (-1.40%)	36.77% (-0.66%)	43.05% (-0.49%)	**42.34% (+0.09%)**	34.13% (-0.46%)	41.39% (-1.71%)
	UNG	62.41% (-2.98%)	42.63% (-1.18%)	44.39% (-1.49%)	53.31% (-4.13%)	36.59% (-0.84%)	41.22% (-2.32%)	41.00% (-1.34%)	33.70% (-0.43%)	41.43% (-1.68%)
CEA 100% Accuracy	MSA	99.07% **(-0.93%)**	98.39% **(-1.61%)**	72.50% **(-27.50%)**	-	-	-	-	-	-
	CEA	100% (n=6336)	100% (n=4224)	100% (n=4495)	100% (n=7698)	100% (n=5015)	100% (n=5834)	100% (n=2847)	100% (n=2331)	100% (n=2238)
	SEA	94.27% (-5.73%)	98.39% **(-1.61%)**	70.43% (-29.57%)	92.61% **(-7.39%)**	91.74% **(-8.26%)**	67.95% **(-32.05%)**	93.64% **(-6.36%)**	94.85% **(-5.15%)**	65.19% **(-34.81%)**
	UNG	90.48% (-9.52%)	91.41% (-8.59%)	68.12% (-31.88%)	85.80% (-14.20%)	88.91% (-11.09%)	62.82% (-37.18%)	87.64% (-12.36%)	93.31% (-6.69%)	64.34% (-35.66%)
SEA 100% Accuracy	MSA	95.95% **(-4.05%)**	94.22% **(-5.78%)**	71.79% **(-28.21%)**	-	-	-	-	-	-
	CEA	95.91% (-4.09%)	94.58% (-5.42%)	72.15% (-27.85%)	94.91% **(-5.09%)**	93.38% **(-6.62%)**	68.71% **(-31.29%)**	93.45% **(-6.55%)**	96.13% **(-3.87%)**	67.89% **(-32.11%)**
	SEA	100% n=(6228)	100% n=(4169)	100% (n=4388)	100% (n=7511)	100% (n=4927)	100% (n=5769)	100% (n=2853)	100% (n=2300)	100% (n=2149)
	UNG	92.68% (-7.32%)	93.84% (-6.16%)	68.69% (-31.31%)	88.63% (-11.37%)	92.10% (-7.9%)	64.21% (-35.79%)	88.78% (-11.22%)	94.91% (-5.09%)	65.57% (-34.43%)

Table 1: **Top:** Overall accuracy on single grammatical feature minimal pairs. **Middle:** Accuracy on the corresponding MSA, SEA, and ungrammatical (UG) variants restricted to items answered correctly in the CEA condition (100% CEA subset). **Bottom:** Accuracy on the corresponding MSA, SEA, and ungrammatical (UG) variants restricted to items answered correctly in the SEA condition (100% SEA subset). Unperturbed items are excluded from the overall evaluation. Random-chance accuracy is 29% for ArabicMMLU and AraDiCE-ArabicMMLU-EGY, and 25% for DialectalArabicMMLU. Bold indicates the highest accuracy within each model and benchmark column. Values in parentheses indicate the absolute percentage-point difference relative to the corresponding CEA-correct or SEA-correct baseline performance.

at a similarly small margin (between -0.5% and -5%). However, it is important to note that finding downstream performance degradation overall (though marginal) in favor of CEA is still noteworthy given that we focus on a single grammatical perturbation at a time, and do not include any syntactic, lexical or orthographic changes in the perturbations. This could indicate that more SEA linguistic dialect markers could contribute in further performance degradation in downstream performance when the full dialect is used. Another possible reason for the marginal difference is that benchmark performance is generally quite low across most models, peaking at around 66% in the best case. Thus, it may be that many cases in which the model answers a question in SEA incorrectly can be attributed to general benchmark difficulty, and not difficulty specific to understanding the dialect. We therefore follow Pan et al. (2025) in also examining the difference in accuracy on the questions which the model got correct in the preferred dialect (in this case, CEA). In other words, if the model got the question correct in the preferred dialect, how likely is the model to get that same question wrong if SEA is used instead and vice versa?

Single Feature Accuracy - CEA/SEA Correct Only Examining only the questions which the model got correct in CEA, the difference between CEA and SEA becomes much more severe. NileChat-4B in particular exhibits an extremely strong effect, with performance dropping nearly 35% when the question is posed in SEA. Llama-3.1-8B by comparison shows the smallest differences almost across the board, which is consistent with its comparative lack of upstream bias observed in Experiment 1. However, these differences are certainly not negligible, with performance degrading up to 8.3% on questions that were answered correctly in CEA. Lastly, Jais-2-8B exists at somewhat of a midpoint, sometimes patterning closely to Llama-3.1-8B and other times showing a stronger bias, but still far less than NileChat-4B. This shows that having an extremely strong upstream preference does not necessarily lead to a proportional impact on downstream performance.

To ensure the robustness of our results, we extend Pan et al. (2025)'s approach looking at prestigious-correct only items to also examining items which the model got correct in the non-preferred dialect (SEA) as well. Examining both

		ARABIC MMLU MSA (2024) % (n=2194)			ARDICE MMLU EGY (2025) % (n=3200)			DIALECTALARABICMMLU (2026) % (n=1844)		
		JAIS 2.0 (8B)	LLAMA 3.0 (8B)	NileChat (4B)	JAIS 2.0 (8B)	LLAMA 3.0 (8B)	NileChat (4B)	JAIS 2.0 (8B)	LLAMA 3.0 (8B)	NileChat (4B)
Overall Accuracy	MSA	**66.00% (+1.55%)**	**45.49% (+6.84%)**	**45.76% (+2.55%)**	-	-	-	-	-	-
	CEA	64.45%	38.65%	43.21%	**57.19%**	**37.22%**	**43.63%**	**43.93%**	**36.44%**	**44.31%**
	SEA	61.17% (-3.28%)	36.65% (-2.01%)	41.29% (-1.91%)	53.25% (-3.94%)	36.44% (-0.78%)	40.03% (-3.59%)	42.46% (-1.46%)	35.03% (-1.41%)	41.59% (-2.71%)
	UNG	55.88% (-8.57%)	34.78% (-3.87%)	37.60% (-5.61%)	48.59% (-8.59%)	36.47% (-0.75%)	38.06% (-5.56%)	39.05% (-4.88%)	35.79% (-0.65%)	37.47% (-6.83%)
CEA 100% Accuracy	MSA	96.68% **(-3.32%)**	91.86% **(-8.14%)**	71.94% **(-28.06%)**	-	-	-	-	-	-
	CEA	100% n=(1647)	100% n=(1003)	100% (n=1100)	100% (n=2223)	100% (n=1450)	100% (n=1696)	100% (n=1259)	100% (n=1053)	100% (n=1270)
	SEA	88.47% (-11.53%)	85.26% (-14.74%)	64.14% (-35.86%)	86.17% **(-13.83%)**	87.49% **(-12.51%)**	64.04% **(-35.96%)**	85.31% **(-14.69%)**	87.80% **(-12.20%)**	64.87% **(-35.13%)**
	UNG	79.42% (-20.58%)	80.07% (-19.93%)	56.65% (-43.35%)	74.10% (-25.90%)	83.63% (-16.37%)	58.81% (-41.19%)	74.69% (-25.31%)	85.57% (-14.43%)	55.57% (-44.43%)
SEA 100% Accuracy	MSA	92.66% **(-7.34%)**	86.36% (-13.64%)	69.61% (-30.39%)	-	-	-	-	-	-
	CEA	92.65% (-7.35%)	89.93% **(-10.07%)**	66.67% **(-33.33%)**	92.36% **(-7.64%)**	89.07% **(-10.93%)**	68.93% **(-31.07%)**	88.21% **(-11.79%)**	91.13% **(-8.87%)**	69.30% **(-30.7%)**
	SEA	100% n=(1566)	100% n=(953)	100% (n=1053)	100% (n=2055)	100% (n=1427)	100% (n=1564)	100% (n=1204)	100% (n=1015)	100% (n=1176)
	UNG	84.22% (-15.78%)	84.78% (-15.22%)	58.31% (-41.69%)	80.92% (-19.08%)	87.74% (-12.26%)	58.89% (-41.11%)	78.49% (-21.51%)	89.06% (-10.94%)	57.99% (-42.01%)

Table 2: **Top:** Overall accuracy on 2+ mixed grammatical feature minimal pairs. **Middle:** Accuracy on the corresponding MSA, SEA, and ungrammatical (UG) variants restricted to items answered correctly in the CEA condition (100% CEA subset). **Bottom:** Accuracy on the corresponding MSA, SEA, and ungrammatical (UG) variants restricted to items answered correctly in the SEA condition (100% SEA subset). Unperturbed items are excluded from the overall evaluation. Random-chance accuracy is 29% for ArabicMMLU and AraDiCE-ArabicMMLU-EGY, and 25% for DialectalArabicMMLU. Bold indicates the highest accuracy within each model and benchmark column. Values in parentheses indicate the absolute percentage-point difference relative to the corresponding CEA-correct and SEA-correct performance.

CEA-correct and SEA-correct items ensures that our analysis does not presuppose that there are no items that the model got correct in CEA but not SEA. We observe in Table 1 that the absolute magnitudes of the number of questions got correct on each benchmark is quite similar between both CEA and SEA. It will therefore be useful to validate the systematicity of the downstream bias with statistical means, since it not clear from the counts/proportions alone that the bias is significant. We will return to this point after reviewing the mixed-feature results.

Mixed-Feature Accuracy Turning to the mixed feature conditions (Table 2), we observe that many of the same patterns hold, but are more pronounced. Both the overall and CEA-correct accuracies are lower in the mixed feature condition than in the single feature condition across all models and benchmarks. Therefore, as more features associated with SEA are introduced into the prompt, the more likely the model is to get the question wrong. This could be for a few different reasons: one reason is simply that the more dialectal features are introduced into the prompt, the more probable it is that the model will encounter a feature with which it is less familiar. Another possibility is that introducing more features makes it more likely that a feature which materially affects the question/answers is changed. For example, changing the plurality of a single noun in the question may not affect the model's ability to recognize the overall meaning of the prompt, whereas replacing a negation marker with one which is no longer comprehensible to the model may impact model performance significantly.

Which models have a significant bias against SEA in downstream performance? We have so far demonstrated that across nearly all models and benchmarks, questions posed to the model in CEA are more likely to be answered correctly than those in SEA. However, the margin of difference for many models and benchmarks is marginal, sometimes within one percent. Moreover, the CEA-correct patterns mirror that of SEA-correct patterns, and therefore, how do we know this is attributed to dialect? To determine which models have significant bias against SEA in downstream performance, we perform McNemar's test between each pair of dialects, treating each question as a sample and whether the question was answered correctly

by the model as the response variable (Pan et al., 2025). We perform the paired tests independently for each model, benchmark, and number of features changed (single vs. mixed). If the result of the test is significant for that pair of dialects, this substantiates that the apparent presence of bias found in the accuracy score differential is not due to chance.

The results show that across nearly all benchmarks, Jais-2-8B exhibits the most consistent bias, with the proportion of correct answers in CEA being significantly higher than those in SEA. The only exception is in the DialectalArabicMMLU benchmark, in which it only becomes significant after three or more SEA features are present. NileChat-4B exhibits a significant bias in favor of CEA on DialectalArabicMMLU, as well as a significant bias in the case where two or more SEA features are present, yet the effect is no longer significant when three or more features are present. In all other benchmarks, the presence of SEA features is not significantly correlated with lower performance on the benchmark. Llama-3.1-8B exhibits the least significant downstream bias, with SEA features resulting in lower benchmark accuracy only in the mixed feature case for ArabicMMLU and in the single feature case for AraDiCE-ArabicMMLU-EGY. Interestingly, in both cases the effect becomes no longer significant once more SEA features are introduced (three or more features for ArabicMMLU, two or more features for AraDiCE-ArabicMMLU-EGY). These results show that the effect of upstream bias on downstream performance varies substantially by model, although no model is completely unanimous across conditions in preferring CEA over SEA or exhibiting no preference.

6 Discussion

The experiments in this paper have shown that the lack of SEA representation in text corpora used to train LLMs leads these models to view sentences with SEA features as less probable than their prestigious CEA counterparts, and that this upstream bias can lead to further impact downstream. It is therefore not the case that we can purely rely on the model's ability to generalize from massive amounts of text in ensuring that the model is equitable towards all Egyptian Arabic speakers.

This demonstration of bias leads us to assert the importance of not viewing dialect too reductively. Although ongoing efforts in Arabic NLP have aimed to face the diglossic nature of the language head-on, a lack of attention to the underlying makeup of major regional dialects implicitly favors the prestigious, "default" dialect. As was mentioned, this leads to an unfortunate feedback loop: the less that existing language technologies support SEA or other non-prestigious sub-dialects, the more likely SEA speakers are to avoid using their native dialect in digital spaces, resulting in less data on which to train and improve the model for that population.

7 Conclusion and Future Work

Given the importance of accounting for sub-dialectal variation, the next step is identifying the existing varieties within each Arabic speaking country, the degree of variation, and whether this variation exists in the upstream representations and affects the downstream performance of LLMs for this dialect. This may require further efforts beyond available digital data, as some of these non-prestigious dialectal speakers could prefer not to position themselves as speakers of these marginalized dialects online, and accordingly we lack the necessary representative data. The present study has focused on dialectal features in an aggregate sense, without particular consideration for exactly which features drive upstream and downstream bias. Future work should aim to establish a causal relationship between individual features and model bias in order to develop a more comprehensive characterization of the problem.

Limitations

The selected grammatical features distinguishing SEA and CEA were drawn from dialect surveys conducted in the 1980s (Behnstedt and Woidich, 1985; Khalafallah, 1969). As these descriptions were collected several decades ago, it is possible that some reported SEA specific features have changed in frequency or fallen out of use over time. To mitigate this limitation, we cross-referenced all selected features against a recently collected spoken Sa'idi Arabic corpus (Mohamed Eida and Habash, 2025) and adjusted the feature set to reflect their observed frequency and usage. We then used regular expressions to automatically generate the MSA, CEA, SEA, and ungrammatical variants by replacing the selected grammatical features, followed by manual validation of the resulting items. Although every generated item was reviewed, some unintended substitutions or annotation errors may

have remained despite our best efforts.

Another limitation of our work is that the models used in both experiments are open source models varying between 350M-13B parameters which are low in size in comparison to larger proprietary models. We understand that the performance on these larger proprietary models perform much better in multiple choice items, however, we were restricted to open source low-moderate sized models due to budgetary and computational constraints.

References

Zaki N Abdel-Malek. 1972. *The closed-list classes of colloquial Egyptian Arabic*. Mouton & Co. N.V., Publishers, The Hauge.

AI@Meta. 2024. Llama 3 model card.

Osamh M Aldosaree. 2016. *Language attitudes toward Saudi dialects*. California State University, Long Beach.

Malik H. Altakrori, Nizar Habash, Teresa Lynn, Younes Samih, Abed Alhakim Freihat, Kirill Chirkunov, Muhammed AbuOdeh, Radu Florian, Preslav Nakov, and Alham Fikri Aji. 2026. Dialectalarabicmmlu: Benchmarking dialectal capabilities in arabic and multilingual language models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 3199–3219, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.

El-Said Badawi. 1973. *Mustawayat al-Arabiyah al-muasirah fī Misr : bahth fi alaqat al-lughah bi-al-hadarah*. Dār al-Mārif, Cairo.

Reem Bassiouney. 2017. *Identity and dialect performance: A study of communities and dialects*. Routledge.

Reem Bassiouney. 2018. Constructing the stereotype: Indexes and performance of a stigmatised local dialect in Egypt. *Multilingua*, 37(3):225–253.

Peter Behnstedt and Manfred Woidich. 1985. Die ägyptisch-Arabischen dialekte. *Tübinger Atlas des Vorderen Orients/Beihefte/B*, 50.

Peter Behnstedt and Manfred Woidich. 2013. Arabic Dialectology. In *The Oxford Handbook of Arabic Linguistics*. Oxford University Press.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744, Singapore. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and 1 others. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Minh Duc Bui, Carolin Holtermann, Valentin Hofmann, Anne Lauscher, and Katharina von der Wense. 2025. Large language models discriminate against speakers of German dialects. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8223–8251, Suzhou, China. Association for Computational Linguistics.

Ibrahim Eltouhamy. 2015. Language attitudes towards dialects of arabic in egypt. Master's thesis, American Univeristy in Cairo.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Howard Giles. 1970. Evaluative reactions to accents. *Educational review*, 22(3):211–227.

Atiqa Hachimi. 2015. "good arabic, bad arabic" mapping language ideologies in the arabic-speaking world. *Zeitschrift für arabische Linguistik*, (61):35–70.

Niloofar Haeri. 1997. *The Sociolinguistic Market of Cairo: Gender, Class, and Education*, 1st edition. Routledge, London.

Muhammad Furquan Hassan, Faiza Khan Khattak, and Laleh Seyyed-Kalantari. 2025. Dialectic preference bias in large language models. *Proceedings of the AAAI Symposium Series*, 5(1):365–369.

Abdelghany A Khalafallah. 1969. *A descriptive grammar of Saidi Egyptian colloquial Arabic*, volume 32. Walter de Gruyter GmbH & Co KG.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

William Labov. 2006. *The social stratification of English in New York city*. Cambridge University Press.

Wallace E Lambert, Richard C Hodgson, Robert C Gardner, and Samuel Fillenbaum. 1960. Evaluational reactions to spoken languages. *The journal of abnormal and social psychology*, 60(1):44.

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. *Proceedings of SCiL*.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342, Vienna, Austria. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Mai Mohamed Eida and Nizar Habash. 2025. Beyond Cairo: Sa'idi Egyptian Arabic literary corpus construction and analysis. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 292–304, Albuquerque, USA. Association for Computational Linguistics.

Mai Mohamed Eida, Mayar Nassar, and Jonathan Dunn. 2024. How well do tweets represent sub-dialects of Egyptian Arabic? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 41–55, Mexico City, Mexico. Association for Computational Linguistics.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Eileen Pan, Anna Seo Gyeong Choi, Maartje Ter Hoeve, Skyler Seto, and Allison Koenecke. 2025. Analyzing dialectical biases in LLMs for knowledge and reasoning benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20882–20893, Suzhou, China. Association for Computational Linguistics.

Dennis R Preston. 1993. Folk dialectology. *American dialect research*, pages 333–378.

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2025. Social bias in large language models for Bangla: An empirical study on gender and religious bias. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 204–218, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, and 3 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Guokan Shang, Hadi Abdine, Ahmad Chamma, Amr Mohamed, Mohamed Anwar, Abdelaziz Bounhar, Omar El Herraoui, Preslav Nakov, Michalis Vazirgiannis, and Eric P. Xing. 2025. Nile-chat: Egyptian language models for Arabic and Latin scripts. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 306–322, Suzhou, China. Association for Computational Linguistics.

Kees Versteegh, Mushira Eid, Alaa Elgibali, Manfred Woidich, and Andrzej Zaborski, editors. 2006–2009. *Encyclopedia of Arabic Language and Linguistics*. Brill, Leiden.

A Probability Distributions

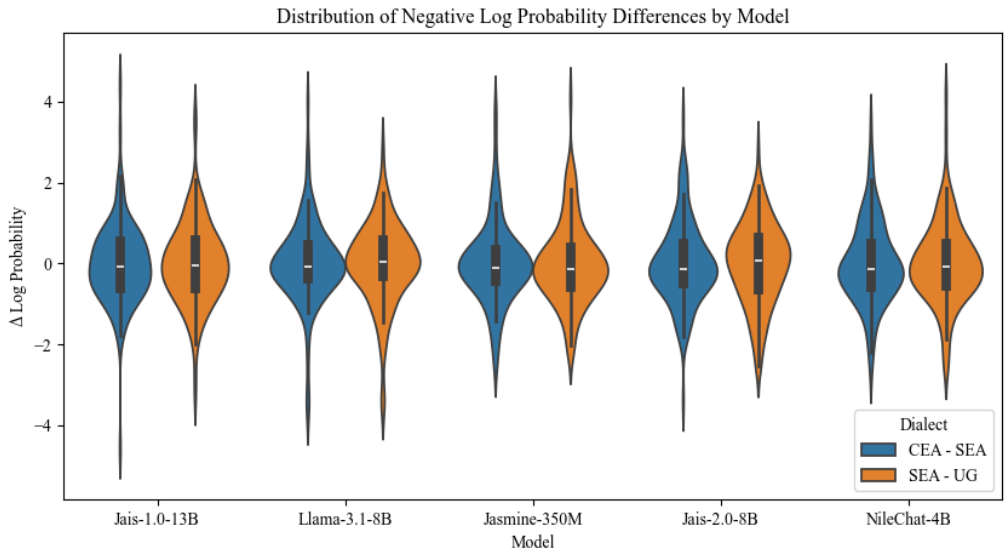

Figure 4: Distribution of differences in negative log probabilities between Cairene Egyptian Arabic (CEA) and Sa'idi Egyptian Arabic (SEA), and between SEA and the ungrammatical baseline (UG) across five language models (Llama-3.1-8B, Jasmine-350M, and Jais-1-13B, Jais-2-8B, and NileChat-4B). Positive values indicate greater preference for the first variety in each comparison (CEA over SEA and SEA over UG), while negative values indicate greater preference for the second.

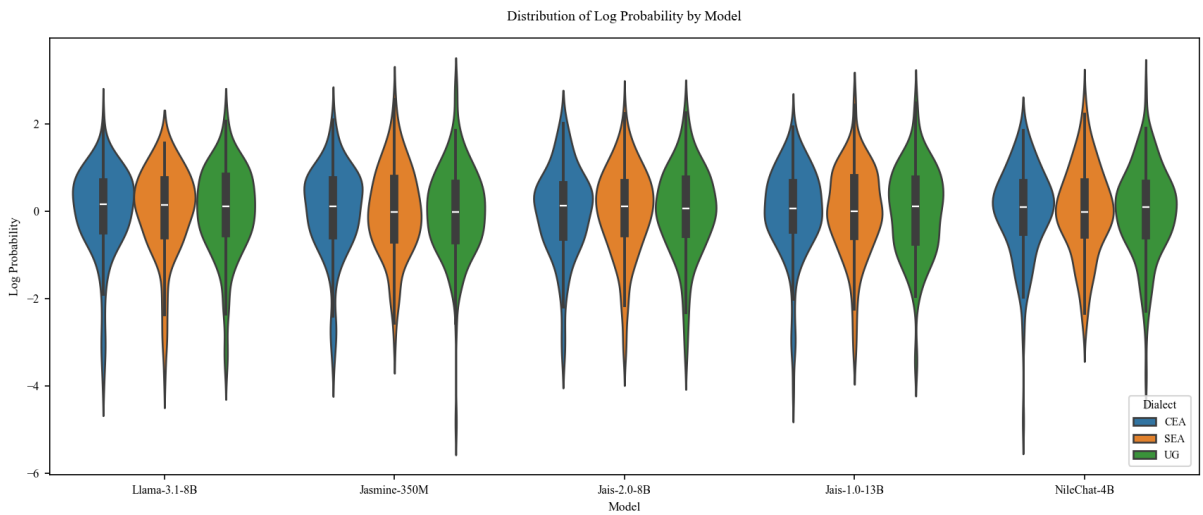

Figure 5: Distribution of negative log probabilities for Cairene Egyptian Arabic (CEA), Sa'idi Egyptian Arabic (SEA), and ungrammatical inputs across five language models (Llama-3.1-8B, Jasmine-350M, and Jais-1-13B, Jais-2-8B, and NileChat-4B). Higher (positive) values indicate greater model preference.

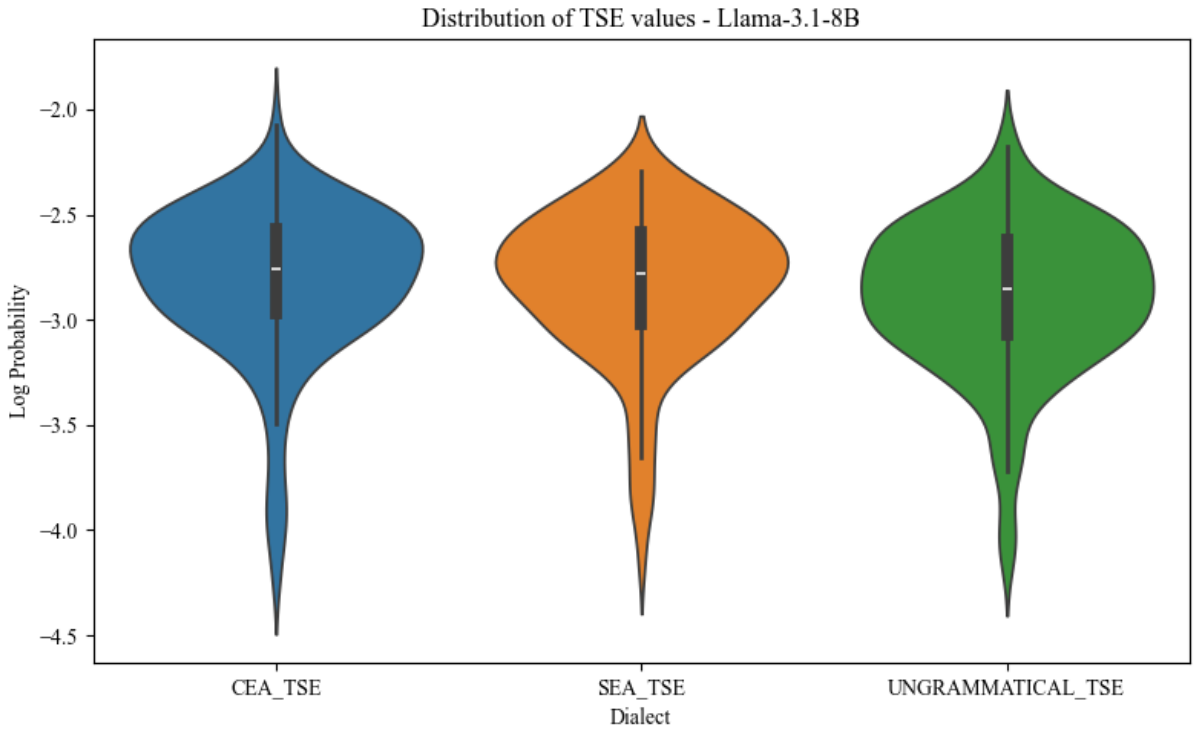

Figure 6: Distribution of token-sequence log probabilities (TSE) for Cairene Egyptian Arabic (CEA), Sa'idi Egyptian Arabic (SEA), and ungrammatical for Llama-3.2-8B.

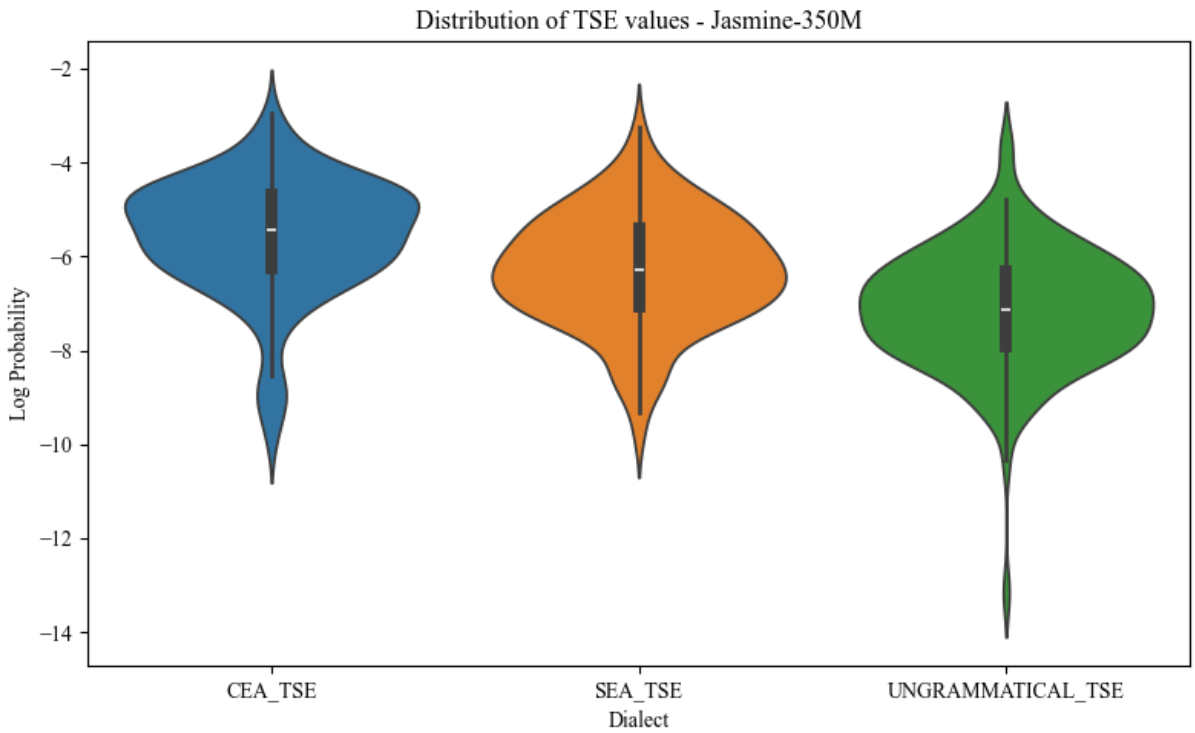

Figure 7: Distribution of token-sequence log probabilities (TSE) for Cairene Egyptian Arabic (CEA), Sa'idi Egyptian Arabic (SEA), and ungrammatical for Jasmine-350M.

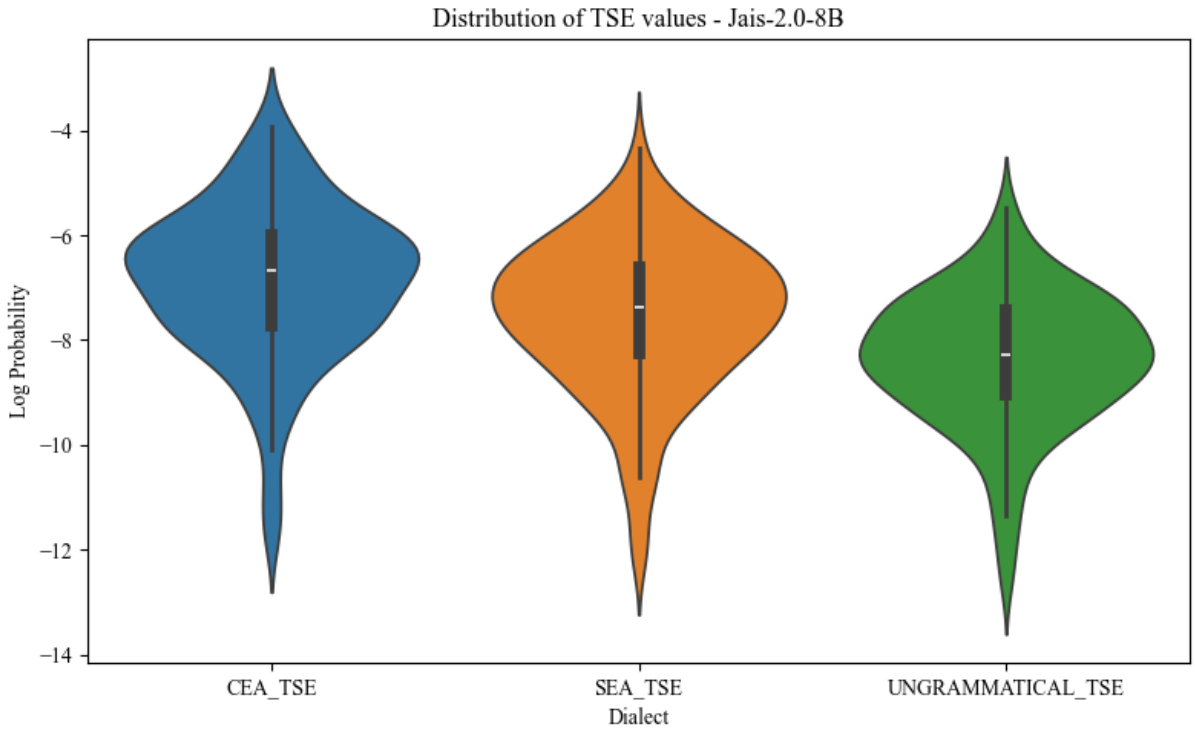

Figure 8: Distribution of token-sequence log probabilities (TSE) for Cairene Egyptian Arabic (CEA), Sa'idi Egyptian Arabic (SEA), and ungrammatical for JAIS-2-8B

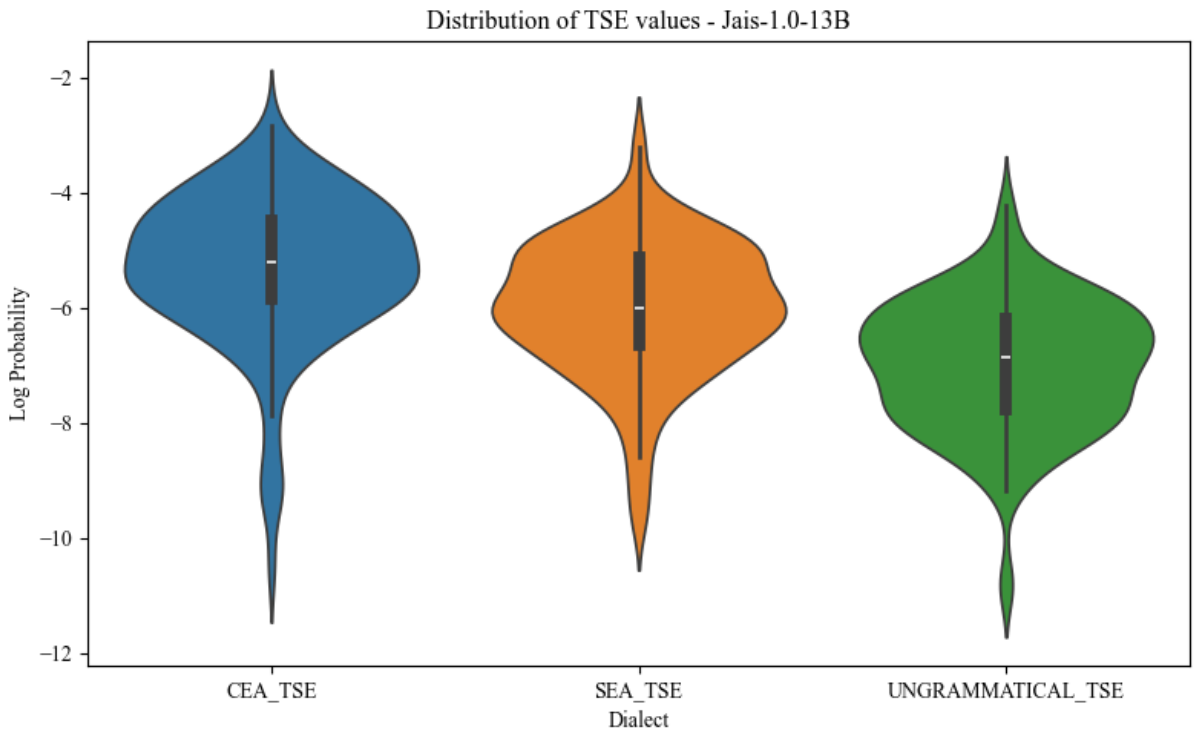

Figure 9: Distribution of token-sequence log probabilities (TSE) for Cairene Egyptian Arabic (CEA), Sa'idi Egyptian Arabic (SEA), and ungrammatical for JAIS-1-13B.

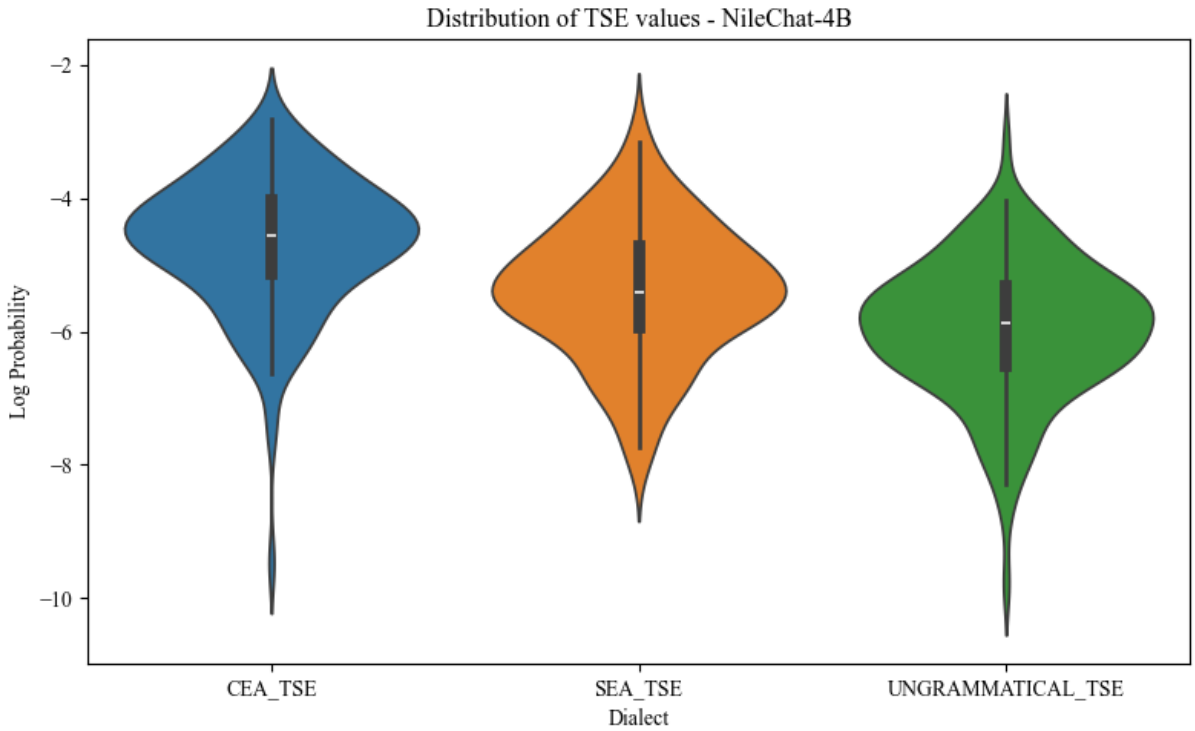

Figure 10: Distribution of token-sequence log probabilities (TSE) for Cairene Egyptian Arabic (CEA), Sa'idi Egyptian Arabic (SEA), and ungrammatical in NileChat-4B.

B Grammatical Features

	# Items	CEA Sample	SEA Sample	Ungrammatical (UG) Sample
Phonetics	41	ليه التأخير ده كله انا حاسس ان الأمور فيها حاجة غلط Why are you so late I feel like something is wrong	ليه التأخير ده كله اني حاسس ان الأمور فيها حاجة غلط Why are you so late I feel like something is wrong	ليه التأخير ده كله لديسلب حاسس ان الأمور فيها حاجة غلط Why are you so late I feel like something is wrong
Morphology	27	في الجوابات الجاية هقولك واحكي لك In my next letters I'll tell you and explain	في الجوابات الجاية عقولك واحكي لك In my next letters I'll tell you and explain	في الجوابات الجاية دقولك واحكي لك In my next letters I'll tell you and explain
Orthography	21	انت مش قلت لى مرة ان امك ميتة Didn't you tell me once your mother had died	انت مش جلت لى مرة ان امك ميتة Didn't you tell me once your mother had died	انت مش سلت لى مرة ان امك ميتة Didn't you tell me once your mother had died
Lexical items	20	وعزيزة امبارح سخنت لكن قايمة الصبح كويسة And Aziza yesterday had a fever but she woke up this morning feeling better	وعزيزة امبارح سخنت لكن قايمة الصبح مليحة And Aziza yesterday had a fever but she woke up this morning feeling better	وعزيزة امبارح سخنت لكن قايمة الصبح ازيلفغ And Aziza yesterday had a fever but she woke up this morning feeling better

Table 3: Samples of CEA–SEA minimal pairs and ungrammatical controls (UG) used to evaluate dialectal preference in LLMs in Experiment 1. This table reports the number of items in each feature category and illustrates the manipulated marked grammatical dialectal features highlighted in red.

Feature	SEA	CEA	MSA	UG	Gloss	ARABIC MMLU MSA	ARADICE MMLU EGY	DIALECTAL ARABIC MMLU
Ad1	دلوق، دلوقت	دلوقتي	الآن	دلوڤ	**Now**	12	11	31
Ad2	واصل	خالص	ابداً	ييلغ	**at all**	5	4	7
Ad3	برا	برة، بره	الخارج	ضصبس	**Outside**	47	20	20
Ad4	برضك، برض	برضه، بردو	أيضاً	وببغلعا	**Also**	5	0	13
Ad5	قوي، جوي	اوي	جداً	ابينتبل	**Very**	25	16	12
Ad6	اهنه، اهنيه	هنا	هنا	بنسليتعبت	**Here**	4	7	26
Ad7	اكده، اكديه	كده، كدا	هكذا	سبسيبيزس	**Like this**	0	48	78
Ad8	لازمن، لازماً	لازم	يجب	ييسظبن	**Have to**	47	393	153
Ad9	بجى	بقى	-	يتزب	**Already**	-	18	27
Dem1	دا	ده	هذا	صيصث	**This**	159	110	381
Introg2	وين	فين	أين	بدشلبس	**Where**	14	17	2
Introg3	ميتى، ميته	امتى	متى	بليظ	**When**	41	33	12
Introg4	كيف	ازاي	كيف	لديظيب	**How**	10	21	2
Prep1	ع	على	على	ز	**On**	2,371	2,120	1,102
Prep2	ف	في	في	ض	**In**	4,593	5,296	1,791
Prep3	كيف	زي	مثل	جش	**Similar to**	89	147	147
Pron1	أني	أنا	أنا	لديسلب	**I**	34	61	20
Pron2	نحنا	احنا	نحن	ضظذر	**We**	0	13	8
RPron3	ال-	اللي	الذي	جشسي	**which**	1,508	3,934	1,547
Neg1	ما + ش	مش	ليس	يوذس	**Negation**	43	408	583
Neg3	ما + شي	ما + ش	لا ، لن	بيسيظبلدبي	**Negation**	511	276	135
Conjun1	لمن	لما	عندما	ضجطب	**when**	79	277	239
Conjun2	لاجل	عشان/علشان	من أجل	جشصس	**Because**	47	46	98
V-Fut1	ع -	ه -	سوف /س -	د	**Future**	45	124	305
					Total	**9,689**	**13,400**	**6,739**

Table 4: SEA, CEA, MSA and UG variants of selected marked grammatical features across each dialect, followed by frequency of benchmark items containing each feature in the ArabicMMLU, AraDiCE-ArabicMMLU-EGY, and DialectalArabicMMLU benchmarks.